IntechOpen

Perturbation Methods with Applications in Science and Engineering

Edited by İlkay Bakırtaş

PERTURBATION METHODS WITH APPLICATIONS IN SCIENCE AND ENGINEERING

Edited by **İlkay Bakırtaş**

Perturbation Methods with Applications in Science and Engineering
http://dx.doi.org/10.5772/intechopen.72260
Edited by İlkay Bakırtaş

Contributors

Kaoru Nakamura, Sadao Higuchi, Toshiharu Ohnuma, Bo Yang, Tao Yu, Hongchun Shu, Pulin Cao, Bahri Sidi Mohammed, Hong Son Hoang, Rémy Baraille, Albert Morozov, Baudel Lara, Arturo Fernández, Lizbeth Morales, Alejandro Altamirano, Albert Reynolds, Cintia Machado, Feng Zhang, Yining Shi, Jia-Ren Yan, Qiu-Run Yu, Jiangnan Li

Notice

First published in London, United Kingdom, 2018 by IntechOpen
eBook (PDF) Published by IntechOpen, 2019
IntechOpen is the global imprint of INTECHOPEN LIMITED, registered in England and Wales, registration number: 11086078, The Shard, 25th floor, 32 London Bridge Street
London, SE19SG – United Kingdom
Printed in Croatia

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Additional hard and PDF copies can be obtained from orders@intechopen.com

Perturbation Methods with Applications in Science and Engineering
Edited by İlkay Bakırtaş

p. cm.

Print ISBN 978-1-78984-255-5

Online ISBN 978-1-78984-256-2

eBook (PDF) ISBN 978-1-83881-668-1

Meet the editor

Dr İlkay Bakırtaş is currently working as associate professor of applied mathematics in Istanbul Technical University (ITU), Department of Mathematics. She received her PhD degree in Mechanics from ITU in 2003. She has completed her post doctoral studies from the University of Colorado at Boulder, USA. She has published 16 research papers in peer reviewed journals in SCI, 4 book chapters, 18 conference proceedings in the fields of perturbation methods, nonlinear wave propagation in arteries, optical solitons and wave collapse in optics and water waves problems. Dr. Bakırtaş is a member of the Scientific Committee of TUMTMK (Turkish National Committee of Theoretical and Applied Mechanics) and she was awarded with "Dr. Serhat Ozyar, Young Scientist Of The Year Award" in 2004 and "Best PhD Dissertation Award" by TUMTMK in 2003.

Contents

Preface

In this book, we aim to present the recent developments and applications of the perturbation theory for treating problems in applied mathematics, physics and engineering. The eight chapters presented in this book are written by 22 authors from 8 countries: Japan, China, Algeria, France, Russia, Mexico, USA and Canada.

Each chapter is independent and self-contained, providing a contemporary overview of the perturbation methods that are used in theoretical and applied sciences. The reference list at the end of each chapter provides the reader a selected list of journal papers, books and conference proceedings. The chapters can be summarized as follows: In the first chapter, a computational technique is developed to predict the piezoelectric properties of materials using the density functional perturbation theory (DFPT). In the next chapter, the development of a sliding-mode perturbation observer-based control scheme for voltage source converter based high voltage direct current systems is described. In the third chapter, a multiplication operation is introduced and via this operation, it is allowed to give the Carleman operator the form of a multiplication operator. In the same chapter, a formal perturbation theory of Carleman operators is also established. The next chapter is devoted to optimal perturbation techniques and various types of optimal perturbation techniques, namely optimal deterministic perturbation theory, optimal stochastic perturbation and simultaneous stochastic perturbation methods are introduced to demonstrate the efficiency of perturbation methods in predictability of dynamical systems that arise in atmospheric and oceanographic sciences. In the fifth chapter, a discussion of the nonlinear parametric systems is presented and the conditions of motions of existence in the resonance zones are put forward. In the sixth chapter, mechanical perturbations strategy is applied at the working electrode during one-step electrodeposition process and the results are compared to the standard one-step electrodeposition. In the next chapter, an approximate analytical solution is constructed for the wellbore pressure via the method of matched asymptotic expansions applied to the one-dimensional saturation convection-dispersion equation. The solutions to this type of nonlinear equations is of great importance in fluid mechanics and especially in petroleum engineering. In the last chapter, a new inhomogeneous scheme based on perturbation methods to solve the solar/infrared radiative transfer (SRT/IRT) problem is developed. This chapter contains significant and applicable information in meteorological sciences.

The book is intended to reach to researchers, scientists and postgraduate students in academia as well as in industry and published as an open access book in order to significantly increase the reach and impact of the information that is contained in the book.

Dr. İlkay Bakırtaş
Istanbul Technical University
Department of Mathematics
Istanbul, Turkey

Chapter 1

Density Functional Perturbation Theory to Predict Piezoelectric Properties

Kaoru Nakamura, Sadao Higuchi and
Toshiharu Ohnuma

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.76827

Abstract

Among the various computational methods in materials science, only first-principles calculation based on the density functional theory has predictability for unknown material. Especially, density functional perturbation theory (DFPT) can effectively calculate the second derivative of the total energy with respect to the atomic displacement. By using DFPT method, we can predict piezoelectric constants, dielectric constants, elastic constants, and phonon dispersion relationship of any given crystal structure. Recently, we established the computational technique to decompose piezoelectric constants into each atomic contribution, which enable us to gain deeper insights to understand the piezoelectricity of material. Therefore, in this chapter, we will introduce the computational framework to predict piezoelectric properties of polar material by means of DFPT and details of decomposition technique of piezoelectric constants. Then, we will show some case studies to predict and discover new piezoelectric material.

Keywords: density functional perturbation theory, ferroelectricity, piezoelectricity, first-principles calculation

1. Introduction

In this chapter, we will introduce how recent computational techniques can successfully predict response properties, represented as piezoelectricity, by means of perturbation method. Piezoelectricity is the polarization change in response to external mechanical force. Inversely, if electrical field is applied to piezoelectric material, mechanical strain is induced (inverse piezoelectric effect). Therefore, piezoelectric materials are widely used as vibrational censors,

surface acoustic wave devices, and actuators. Only the material having no inversion symmetry shows piezoelectricity. For example, **Figure 1** shows schematic illustration of the piezoelectric effect. Positions of positively charged ion (cation) and negatively charged ion (anion) are represented as plus and minus symbols. **Figure 1a** shows the paraelectric phase, where ions are orderly located with inversion symmetry. On the other hand, ions are slightly displaced by δ with respect to those in paraelectric phase, as shown in **Figure 1b**. Such small displacement induces microscopic polarization P_s along the ionic displaced direction.

Because ferroelectric phase is energetically more stable than paraelectric phase under low temperature, P_s is frequently referred as the spontaneous polarization. Above Curie temperature, ferroelectric properties are disappeared since paraelectric phase becomes more stable than ferroelectric one. **Figure 1c** shows the schematic illustration of the principal of piezoelectricity, where external stress (red-colored arrows) increases the ionic displacement and resultant polarization. In this case, external stress increases the spontaneous polarization by $\Delta P_s = P_s' - P_s$. Therefore, piezoelectric constant is defined as the derivative of the spontaneous polarization with respect to the external field. More detailed and comprehensive description of piezoelectricity is reviewed by Martin [1].

First-principles calculation based on density functional theory (DFT [2, 3]) has been widely utilized as the computational method to predict the electronic properties of material under the ground state. Ideally, required information to conduct the first-principles calculation is only the crystal structure, including atomic species and position of periodic/nonperiodic structure unit. The most significant advantage of first-principles calculation is its predictability. Since King-Smith and Vanderbilt showed the theoretical methodology to calculate change in polarization per unit volume ΔP [4], dielectric and piezoelectric properties of wide range of materials in which electronic correlations are not too strong [5–7] have been accurately predicted. The derivative of total energy determines various properties. For example, determined forces, stresses, dipole moment (first-order derivatives), dynamical matrix, elastic constants, dielectric and piezoelectric constants (second-order derivative), nonlinear dielectric susceptibility, phonon–phonon interaction and Grüneisen parameters (third-order derivative), and

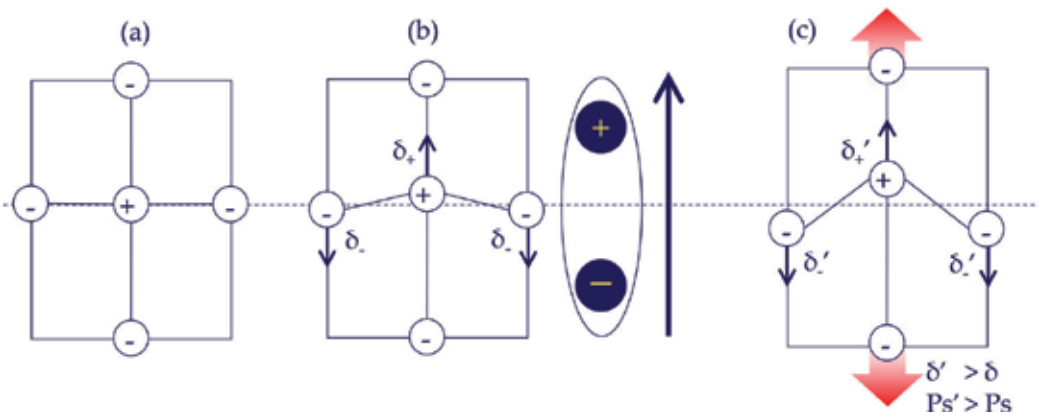

Figure 1. Ionic configuration of (a) paraelectric phase and (b) ferroelectric phase. (c) Ionic displacement according to the external force.

so on. Thus, first-principles calculation has been made use of calculating the perturbed total energy of materials because of its accuracy. Although perturbations were made by hand up to the early 1980s, sophisticated methodology of density functional perturbation theory (DFPT) was proposed in 1987 by Baroni et al. [5]. They showed general formulation of total energy change with respect to atomic displacement and opened the way to efficiently compute the energy derivative with respect to the perturbation [5]. DFPT can compute response properties directly arising from the perturbations of strain, atomic displacement, and electric field by making use of linear response theory [8–11]. Numbers of ferroelectric materials are theoretically investigated on the origin of their ferroelectric properties (including piezoelectricity and dielectric properties) by using DFPT. Because of technological importance, such theoretical researches have been focused on Pb-based perovskite material (e.g., $PbTiO_3$, $PbZrO_3$, and their solid solution [12–15]) because they have excellent piezoelectric properties and are widely applied for actuators. However, due to the restriction of hazardous substance (RoHS) directive, researches on lead-free ferroelectric materials have gathered great attraction. By taking advantage of the predictability of DFPT, various lead-free ferroelectric oxide and nitride materials were theoretically investigated on their piezoelectric properties [16–25]. Moreover, DFPT calculations showed that piezoelectricity can be greatly enhanced by imposing isotropic stress for $PbTiO_3$ [26, 27], uniaxial stress for $SrHfO_3$ [28], uniaxial and biaxial strain for AlN-GaN solid solution alloy [29], and two-dimensional epitaxial strain for doped ZnO [30]. As latterly explained, those enhancements of piezoelectric constant are thought to be closely related to the phase transition. In the next section, we will show the definition of piezoelectric constants within the framework of DFPT.

2. Formulation of piezoelectric constants

Formulation and calculation methodologies to obtain response properties of materials in the framework of DFPT have been developed in a step-by-step manner, because degrees of freedom by perturbations of atomic displacement, homogeneous electric fields, and strain are often strongly coupled. For example, piezoelectricity affects elastic and dielectric properties. Therefore, special care must be paid for the calculation of coupled properties. In 2005, Hamann et al. demonstrated that elastic and piezoelectric tensors can be efficiently calculated by treating homogeneous strain within the framework of DFPT [31]. At the same time, Wu et al. systematically formulated response properties with respect to displacement, strain, and electric fields [32]. In this section, we will briefly introduce how piezoelectric properties are formulated in the framework of DFPT. In each formulation, Einstein implied-sum notation is used. Cartesian directions $\{x, y, z\}$ are represented as α and β. Subscription of j and $k = 1, \ldots, 6$ is the standard Voigt notation (represents directions of xx, yy, zz, yz, zx, and xy). The subscripts m and n are the degrees of freedom in the cell. They range from 1 to $3i$, where i is the number of irreducible atoms because each atom has three degree of freedom along x, y, and z directions.

Total energy of material under perturbation of atomic displacement u, electric field σ, and strain η, $E(u,\sigma,\eta)$, is defined as follows:

$$E(u, \sigma, \eta) = \frac{1}{\Omega_0}[E^0 - \Omega\sigma \cdot \mathrm{P}] \tag{1}$$

where E^0 is the total energy of material under the ground state, Ω_0 is volume of the unit cell (smallest repeat unit of crystal), Ω is deformed volume of the unit cell, and P is the electric polarization. Following response functional tensor can be obtained by second-order differential of Eq. (1):

$$\text{Force constant matrix}: K_{mn} = \Omega_0 \frac{\partial^2 E}{\partial u_m \partial u_n}\bigg|_{\sigma,\eta} \tag{2}$$

$$\text{Clamped} - \text{ion term of electric susceptibility}: \bar{\chi}_{\alpha\beta} = -\frac{\partial^2 E}{\partial \sigma_\alpha \partial \sigma_\beta}\bigg|_{u,\eta} \tag{3}$$

$$\text{Clamped} - \text{ion term of elastic tensor}: \bar{C}_{jk} = -\frac{\partial^2 E}{\partial \eta_j \partial \eta_k}\bigg|_{u,\sigma} \tag{4}$$

$$\text{Born effective charge tensor}: Z_{m\alpha} = -\Omega_0 \frac{\partial^2 E}{\partial u_m \partial \sigma_\alpha}\bigg|_{\eta} \tag{5}$$

$$\text{Force} - \text{response internal} - \text{strain tensor}: \Lambda_{mj} = -\Omega_0 \frac{\partial^2 E}{\partial u_m \partial \eta_j}\bigg|_{\sigma} \tag{6}$$

$$\text{Clamped} - \text{ion piezoelectric tensor}: \bar{e}_{\alpha j} = \frac{\partial^2 E}{\partial \sigma_\alpha \partial \eta_j}\bigg|_{u} \tag{7}$$

Clamped-ion term is a frozen quantity, which indicates that atomic coordinates are not allowed to relax as the homogeneous electric field or strain. Therefore, dynamical term should be added into the clamped-ion term in order to obtain proper response properties.

Simplest and physically well-understandable piezoelectric constant can be expressed as follows:

$$e_{\alpha j} = \frac{\partial P_\alpha}{\partial \eta_j} \tag{8}$$

In this expression, it is easily understood that piezoelectric e constant $e_{\alpha j}$ is a measure of the change in polarization induced by the external strain. As the atomic positions are changed according to the strain, change of the polarization includes both electronic contribution (clamped-ion term) and dynamical contribution (internal-strain term). The internal-strain term of piezoelectric constant is represented as follows:

$$\hat{e}_{\alpha j} = \frac{1}{\Omega_0} Z_{m\alpha} (K^{-1})_{mn} \Lambda_{nj} \tag{9}$$

Thus, proper piezoelectric constant can be obtained by Eqs. (7) and (9):

$$e_{\alpha j} = \left.\frac{\partial^2 E}{\partial \sigma_\alpha \partial \eta_j}\right|_u + \frac{1}{\Omega_0} Z_{m\alpha} (K^{-1})_{mn} \Lambda_{nj} \tag{10}$$

Here, the first and second terms on the right-hand side in Eq. (10) are the clamped-ion term and internal-strain term, respectively. The former shows the electronic contribution ignoring the atomic relaxation effect, and the latter shows the ionic contribution including the response of the atomic displacement to the strain. The Born effective charge $Z_{m\alpha}$, force-constant matrix K_{mn}, and internal-strain tensor Λ_{nj} are the second derivatives of the energy with respect to the displacement and electric field, pairs of displacements, and displacement and strain, respectively. The internal-strain term of the piezoelectric stress constants can be further decomposed into the individual atomic contributions when the above second-derivative tensors are fully obtained.

On the other hand, the internal-strain term of the piezoelectric stress constant $e_{\alpha j}$ is frequently described by the following equation, using the Born effective charge $Z_{\alpha\beta}$ and displacement u_β of each atom in the calculation cell:

$$\hat{e}_{\alpha j} = Z_{\alpha\beta} \frac{\partial u_\beta}{\partial \eta_j} \tag{11}$$

where $\partial u_\beta / \partial \eta_j$ shows the response of the first-order atomic displacement to the first-order strain. In this expression, the meaning of the piezoelectric stress constant, i.e., e_j is a measure of the change in polarization induced by the external strain, is much more visible than in Eq. (9). In the DFPT formalism, $\partial u_\beta / \partial \eta_j$ is implicitly calculated as a displacement-response internal-strain tensor Γ as follows [32]:

$$\Gamma_{nj} = \Lambda_{mj} (K^{-1})_{mn} \tag{12}$$

Because the subscript n in Γ_{nj} indicates the degrees of freedom, Γ_{nj} can be decomposed into the individual atomic components, which also enables to calculate individual contribution of each atom for total piezoelectric constant.

Here, piezoelectric e constant defined as Eq. (9) is frequently referred as "piezoelectric strain constant." On the other hand, it is much more natural and easy to control the stress (electric field) than to control the strain in any case. In this case, the piezoelectric strain constant $d_{\alpha j}$ is usually measured. It can be obtained from piezoelectric strain constant $e_{\alpha j}$ using the following relation:

$$d_{\alpha j} = s_{jk} e_{\alpha k} \tag{13}$$

where s_{jk} is the elastic compliance, which is given by the inverse matrix of the elastic constants C_{jk}.

Those formulations are implemented in specific first-principles simulation packages such as ABINIT [33] and Vienna ab initio simulation package (VASP) [34], and piezoelectric constants can be calculated on a daily basis. From the next section, we will show how DFPT calculation precisely gives piezoelectric properties of ferroelectric materials.

3. Introduction of target material

In this chapter, we selected $LiNbO_3$ as a target material to show the predictability of DFPT calculation. $LiNbO_3$ is one of ferroelectric materials and widely used as surface acoustic wave (SAW) and optical waveguide elements. Crystal structure of $LiNbO_3$, which belongs to the space group of *R3c*, is frequently referred as "strained perovskite structure." Schematic illustrations of crystal structure of ABO_3 perovskite and $LiNbO_3$ are shown in **Figure 2**.

Crystal structures shown in the present chapter was visualized by using VESTA software [35]. Curie temperature of $LiNbO_3$ is quite high and ranges from 1140 [36] to 1210°C dependent on the quality of sample (variation of Li/Nb relation can shift Curie temperature [37]). Below Curie temperature, ferroelectric phase with *R3c* symmetry (crystal structure can be classified into 230 types of space group according to the symmetry group) shown in **Figure 2a** is stable. Paraelectric phase with $R\bar{3}c$ symmetry, shown in **Figure 2b**, becomes stable above Curie temperature. In the paraelectric phase, it can be seen that both Li and oxygen is positioned with

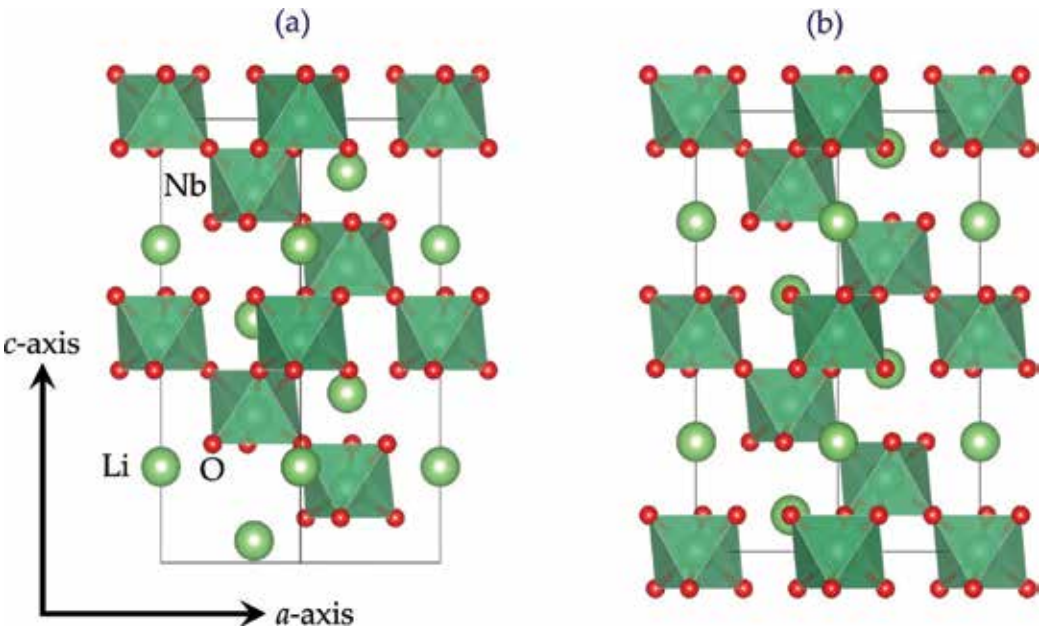

Figure 2. Crystal structures of (a) ferroelectric phase with *R3c* symmetry and (b) paraelectric phase with $R\bar{3}c$ symmetry $LiNbO_3$. Yellow green-, green-, and red-colored balls represent Li, Nb, and oxygen atoms, respectively. Bonding structures between Nb and surrounding oxygen atoms are represented as green-colored polyhedron. Two orthogonal crystallographic directions are shown as both *a*- and *c*-axes.

the same height along *c*-axis, and the position of Nb is just the center between two oxygen layers along *c*-axis. On the other hand, both Li and Nb are shifted in ferroelectric *R3c* phase along downward direction of *c*-axis with respect to those in paraelectric $R\bar{3}c$ phase.

Due to the different bonding nature between Li-O and Nb-O, atomic positions of Li and Nb are off-centered within oxygen layers along *c*-axis. This structural characteristic is the ferroelectric nature of $LiNbO_3$. One of the notable properties of $LiNbO_3$ is its high-curie temperature (~1400 K). However, piezoelectric properties of $LiNbO_3$ are not so much superior as compared with Pb-based perovskites. Crystal structure of piezoelectric ABO_3 perovskite is based on the cubic structure (of *Pm3m* symmetry), shown in **Figure 3a**.

Cubic lattice is symmetric and usually high-temperature phase, same as $LiNbO_3$. The "strained perovskite structure" expression for $LiNbO_3$ means that LiO_6 and NbO_6 polyhedron are largely rotated with respect to the cubic perovskite structure. However, because of the simple atomic configuration of cubic structure, atoms can be displaced along various directions and change crystalline symmetry as shown in **Figure 3a**. Crystalline lattice is vibrated (referred as phonon) under finite temperature. Some lattice vibrations along specific directions are unstable. This specific phonon is called as soft mode with imaginary frequency. In such case, atoms are displaced along unstable phonon mode to lower the total energy. For example, cooperative atomic displacement along [001] direction shown in **Figure 3b** (referred as Γ_{15} mode) changes symmetry from cubic to tetragonal (of *P4mm* symmetry), which leads polarization along [001] direction. Thus, polarization direction of perovskite is not restricted and allowed to be changed. This characteristic rotational polarization direction is favorable for piezoelectricity because grains in polycrystalline material are oriented along various directions. Thus, careful controlling of crystal structure is essential to obtain superior piezoelectric properties.

The most convenient way to control and drastically change the crystal structure is imposing high pressure. Many compounds have found to be possible to form $LiNbO_3$-type structure under the high-pressure synthesis [38], and some of them were quenchable phase. For example, $LiNbO_3$-

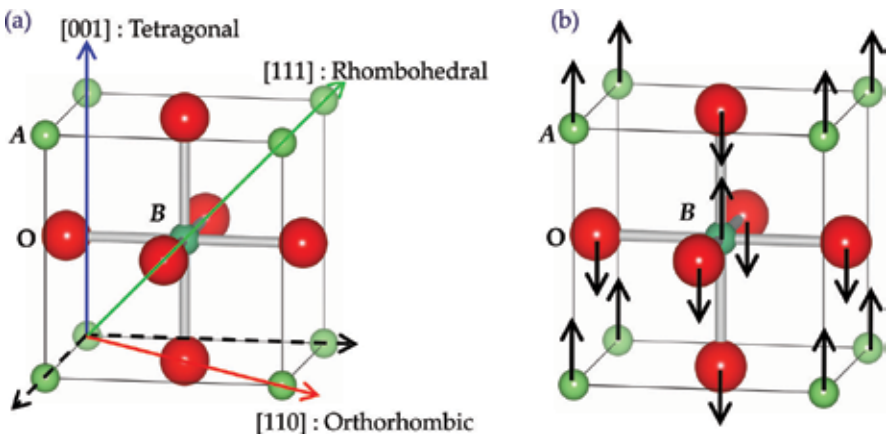

Figure 3. (a) Crystal structure of cubic ABO_3 perovskite and possible polarization directions. (b) Representative unstable vibrational mode of cubic ABO_3 perovskite showing as arrows.

type structured $ZnSbO_3$ was successfully synthesized [39] under high pressure, and improvement of the spontaneous polarization is suggested by enhancement of the covalency of Sn site from first-principles simulation [40]. Moreover, high-pressure synthesized research on $LiNbO_3$-type structure is now extended to more complex compounds such as oxynitrides [41, 42].

The crystal structure of ABO_3 compound is generally determined by the balance between the ionic radius of *A* and *B* element, which is frequently referred as tolerance factor. Due to the small size of the Li ion with respect to the tolerance factor of $LiNbO_3$, this compound cannot form stably the popular perovskite structure under the ambient condition. On the other hand, we predicted the crystal structures of high-pressure phase of $LiNbO_3$ [43], which were not completely elucidated by experimental study [44]. Revealed structures are $NaIO_3$-type structure (*Pnma*) as room temperature high-pressure phase and apatite-like structure ($P6_3/m$) as high-temperature high-pressure phase. It should be noted that the $NaIO_3$-type structure is closely related with the popular $GdFeO_3$-type perovskite structure. The only difference between these structures is that A-site position and B-site position are inter-exchanged. Therefore, there seems to be a possible way to connect the perovskite structure and $LiNbO_3$-type structure.

In our previous theoretical study on high-pressure phase, analysis was mainly concerned with phase transition mechanism only from the viewpoint of subgroup symmetry and energy barrier [43]. It will be instructive to deal with this phase transition phenomenon from the viewpoint of lattice instability as discussed in the field of the ferroelectric instability analysis. In the following section, we will show investigation on the potential piezoelectric properties of $LiNbO_3$ with various hypothetical crystal structures by the method of DFPT, and possible phase transition mechanism will be discussed from the viewpoint of soft mode.

4. Computational methodology

First-principles calculation was performed by using VASP code [34]. Interactions between ion and electron were treated by projector augmented wave (PAW) method [45]. PBEsol functional [46] was used to approximate exchanges and correlate interactions of electrons, which can be used to reproduce the lattice constants of various materials [45]. Precise calculation on the lattice constant is essential to predict piezoelectric properties because they depend on volume of unit cell Ω as shown in Eq. (1). The kinetic energy cutoff for plane waves was set at 500 eV, and the k-point mesh was set at ~0.03/Å intervals to obtain the converged total energy at less than 0.1 meV/atom. Before calculating the piezoelectric constants, atomic positions and cell parameters were optimized until the forces on each atom and cell converged at below 5×10^{-4} eV/Å.

Since VASP does not directly calculate Eq. (12), we added routine to calculate displacement-response internal-strain tensor Γ_{nj} and decompose piezoelectric constants into each atom. The sum of the decomposed piezoelectric constants was confirmed to accurately reproduce the total piezoelectric constants. Careful convergence tests with a higher energy cutoff and denser *k*-point mesh showed that the numerical accuracy of the calculated Γ_{nj} was less than 0.01. It was confirmed that this error does not influence our discussion and conclusion. Moreover, it was confirmed that the values of Γ_{nj} obtained by the DFPT method were consistent with those

calculated by the direct method, in which the strain-displacement relation of each ion was explicitly calculated.

On the basis of cubic *Pm3m* phase, lattice instability analysis was performed by phonon calculation utilizing phonopy code [47]. Force constant matrix shown in Eq. (2) was constructed by DFPT calculation implemented in VASP code combined with supercell approach. Supercell was constructed by using unit cell so that orthogonal three axes of the supercell exceed 10 Å. Note that although supercell is not required in DFPT approach, the present VASP code implements perturbation at the zone center.

5. Calculated piezoelectric properties of $LiNbO_3$

Calculated piezoelectric properties of $LiNbO_3$ in ferroelectric phase are summarized in **Table 1**. Some experimentally measured values are also shown in **Table 1**. All properties are confirmed to be well reproduced by calculation. In a technological importance, 33 components are the most important because *C*-axis of $LiNbO_3$ is polarization direction. Calculated values of e_{33}, C_{33}, and ε_{33} are especially well reproduced. It should be mentioned here that chemical composition of $LiNbO_3$ used for experiment is congruent and includes Li vacancy. On the other hand, calculation was performed by using stoichiometric $LiNbO_3$.

	Calculated value	Experimental value
Piezoelectric stress constant (C/m^2)		
e_{15}	3.73	3.655 ± 0.022 [48], 3.7 [49]
e_{22}	2.51	2.407 ± 0.015 [48], 2.5 [49]
e_{31}	0.21	0.328 ± 0.032 [48], 0.2 [49]
e_{33}	1.69	1.894 ± 0.054 [48], 1.3 [49]
Elastic constant (GPa)		
C_{11}	190.7	198.86 ± 0.033 [48], 203 [49]
C_{12}	58.3	54.67 ± 0.04 [48], 53 [49]
C_{13}	62.4	67.99 ± 0.55 [48], 75 [49]
C_{14}	13.5	7.83 ± 0.02 [48], 9 [49]
C_{33}	220.0	234.18 ± 0.75 [48], 245 [49]
C_{44}	49.2	59.85 ± 0.01 [48], 60 [49]
Dielectric constant		
ε_{11}	40.6	44.9 ± 0.4 [48], 44 [49]
ε_{33}	24.1	26.7 ± 0.3 [48], 29 [49]

Table 1. Piezoelectric constant, elastic constant, and dielectric constant calculated by DFPT and experimentally measured values.

Decomposed e_{33}' (C/m^2)			Born effective charge Z_{33} (e)			Displacement-response internal-strain constant Γ_{33}		
Li	Nb	O	Li	Nb	O	Li	Nb	O
0.1	0.05	0.16	1.03	6.77	−2.6	0.67	−0.05	−0.21

Table 2. Decomposed piezoelectric constants of $LiNbO_3$.

Thus, Li vacancy is considered to have negligible influence on the piezoelectric properties. Decomposed ionic contribution of piezoelectric strain constant e_{33} is summarized in **Table 2**. Although the Born effective charge of Nb is larger than its formal charge +5e, displacement-response internal-strain constant of Nb is negative value. This indicates that piezoelectricity of $LiNbO_3$ is mainly dominated by displacement of Li. Born effective charge indicates a degree of polarization induced by atomic displacement and dominated by the change in the orbital hybridization. Although anomalously large Born effective charge is crucial for superior piezoelectric properties of perovskite ABO_3 materials [50], the present study of decomposition of piezoelectric constant shows that coupling degree between external strain and atomic displacement is also indispensable to understand the piezoelectric properties.

6. Piezoelectric properties of perovskite-$LiNbO_3$

Next, we will show how piezoelectric properties are affected by crystal structure, while chemical composition is kept as $LiNbO_3$. Various hypothetical crystal structures common for perovskite-type structure were constructed, and their energetic stabilities were examined by calculating enthalpy $H = U + PV$ (U is total energy obtained by first-principles calculation, P is external pressure, and V is equilibrium volume under pressure P) as a function of external pressure. Imposing high pressure is most convenient method to modify crystal structure and find unexpected stable phase. The following eight types of phases were considered:

Cubic, *Pm-3 m*; tetragonal, *P4mm*; and rhombohedral, *R-3 m*.

$LiNbO_3$-ferroelectric phase, *R3c*, and $LiNbO_3$-paraelectric phase, *R-3c*.

Orthorhombic, *Amm2* and *Cmmm*, and high-pressure phase, *P63/m.*

where names of space groups are used to distinguish each structure. Crystal structure of each phase is shown in **Figure 4a**. Polyhedra shown in **Figure 4a** correspond to Nb-centered bonding structure of Nb-O bondings. **Figure 4b** shows the enthalpy difference of each phases as a function of external pressure. Here, external pressure is assumed to be isotropic. Standard of enthalpy was set to be the enthalpy of most stable *R3c* phase under ambient condition. At the positive (compressive) pressure region, $P6_3/m$ phase becomes stable above 21 GPa, which is close to the experimental phase transition pressure of 25 GPa [43]. Details of the phase transition behavior under high pressure are theoretically investigated in our previous work [44]. Unfortunately, $P6_3/m$ phase is highly symmetric and shows no piezoelectricity. At the negative (expansive) pressure region, enthalpy difference becomes smaller as there is an increase of negative pressure except for $R\bar{3}c$ and $P6_3/m$ phases.

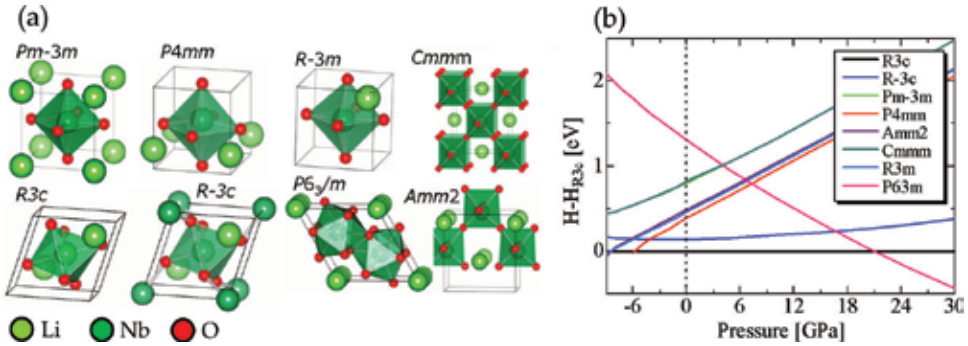

Figure 4. (a) Schematic illustration of eight types of perovskite-structured $LiNbO_3$ and their space groups. (b) Enthalpy differences of each phase measured from the enthalpy of *R3c* phase as a function of pressure.

Imposing negative pressure can be achieved by solid solution with parent phase of larger lattice constant. At −6 GPa, *P4mm* phase becomes stable, while *R3m* and *Amm2* phases become stable at −9 GPa. However, bond breaking occurs in Nb-O bonding above −6 GPa for *P4mm* phase. The same bond breaking occurs in *Amm2* and *R3m* phases at −11GPa and −14 GPa, respectively. Thus, those phase transitions occur just before bond breaking.

Within the eight phases shown in **Figure 4a**, only *P4mm*, *R3m*, *R3c*, and *Amm2* phases show piezoelectricity. Piezoelectric stress constant, elastic constant, and dielectric constant of *P4mm*, *R3m*, and *Amm2* phases are compared with those of *R3c* phase in **Table 3**. Various piezoelectric properties are observed by each phase. Especially for *P4mm* and *Amm2* phases, high e_{33} and relatively low C_{33} values are predicted, which are advantageous for large piezoelectric strain constant d_{33}. On the other hand, *R3m* phase was found to be unstable because following mechanical stability conditions of rhombohedral symmetry:

$$C_{11}+C_{12}>0, C_{33}>0, (C_{11}+C_{12})^{*}C_{33}>2C_{13}, C_{11}-C_{12}>0, C_{44}>0, (C_{11}-C_{12})^{*}C_{44}>2C_{14} \quad (14)$$

are broken because of $C_{44}<0$.

Figure 5a and **b** show piezoelectric properties of *P4mm* phase as a function of pressure and corresponding volume of unit cell. Dotted lines indicate zero pressure states. Piezoelectric stress constant e_{33} of *P4mm* phase shows parabolic behavior and maximum value at zero pressure state. On the other hand, elastic constant C_{33} of *P4mm* phase continuously decreases as volume increases, because orbital hybridization of Nb-O bonding along polarization direction decreases as bond length increases. At the pressure of −6 GPa, C_{33} of *P4mm* phase shows almost zero value. This indicates that Nb-O bonding is broken. Piezoelectric stress constant d_{33} shown in **Figure 5b** increases as volume, because of increase of elastic compliance. Especially at the pressure of −5 GPa just before bond breaking, d_{33} shows maximum value or approximately 1000 pC/N.

This giant piezoelectric constant is almost comparable to that of PZT material [51]. Giant piezoelectric constant is understood as a result of phase instability in morphotropic phase boundary [52]. The same as *P4mm* phase of $LiNbO_3$, we revealed that ZnO also showed anomalously large piezoelectric constant just before phase transition [30].

	R3c	*P4mm*	*R3m*	*Amm2*
Piezoelectric stress constant (C/m^2)				
e_{15}	3. 73	1.14	5.10	0.64
e_{22}	2.51	—	1.19	—
e_{31}	0.21	0.46	0.24	0.80
e_{33}	1.69	3.28	1.92	2.99
Elastic constant (GPa)				
C_{11}	190.7	297.2	203.0	321.8
C_{12}	58.3	48.9	169.0	91.6
C_{13}	62.4	77.7	90.6	92.8
C_{14}	13.5	—	−42.1	—
C_{33}	220.0	157.7	206.6	176.4
C_{44}	49.2	39.5	−29.6	32.5
Dielectric constant				
ε_{11}	40.6	56.2	36.6	28.2
ε_{33}	24.1	12.3	16.2	13.3

Table 3. Piezoelectric constant, elastic constant, and dielectric constant of ***R3c***, ***P4mm***, ***R3m***, and ***Amm2*** phases calculated by DFPT.

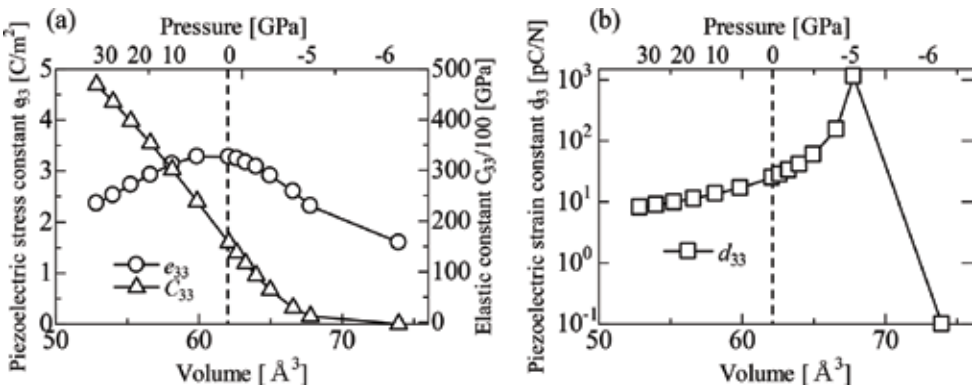

Figure 5. (a) Piezoelectric stress constant e_{33} and elastic constant C_{33} and (b) piezoelectric strain constant d_{33} of *P4mm* phase as a function of pressure and corresponding volume of unit cell.

Finally, we would like to show phase transition path between cubic perovskite structure and $LiNbO_3$ structure. **Figure 6a** shows the energy change of *Pm3m* phase as a function of Li displacement along <001>, <011>, and <111> directions. *Pm3m* phase is paraelectric phase. Because ferroelectricity and piezoelectricity of $LiNbO_3$ are dominated by off-centering and displacement of Li, respectively, phase transition from *Pm3m* phase is also expected to be occurred by

Li displacement. Li displacement along <001>, <011>, and <111> directions induces tetragonal, orthorhombic, and rhombohedral phase transition from cubic phase. **Figure 6a** clearly shows that tetragonal phase transition from *Pm3m* phase to *P4mm* phase is the most energetically advantageous. **Figure 6b** shows the phonon dispersion curve of cubic *Pm3m* phase of $LiNbO_3$. Horizontal axis corresponds to sampling path along high symmetric reciprocal point (*q*-point). Within the whole Brillouin zone of reciprocal space, unstable phonon modes with imaginary phonon frequencies are observed. Here, imaginary phonon frequency is represented as negative value for convenience. Therefore, cubic *Pm3m* phase of $LiNbO_3$ is thermodynamically unstable and considered to show phase transition in accordance with specific phonon mode of imaginary frequency (referred as soft mode). Thus, modulated structures were constructed by imposing atomic displacement along normal modes at each symmetric *q*-points. Modulated structures were structurally relaxed, and their space group and energy change from cubic *P4mm* phase were investigated. Summary of such modulated structures are shown in **Table 4**. At Γ point, tetragonal phase transition along with Γ_{15} soft mode of cubic phase shown in **Figure 3b** shows energy gain of −0.422 eV/formula unit (f.u.). On the other hand, it was found that modulation at R point gives more stable energy gain of −0.682 eV/f.u. In this case, R_{25} soft mode induces phase transition from cubic *Pm3m* phase to $R\bar{3}c$ phase shown in **Figure 2b**.

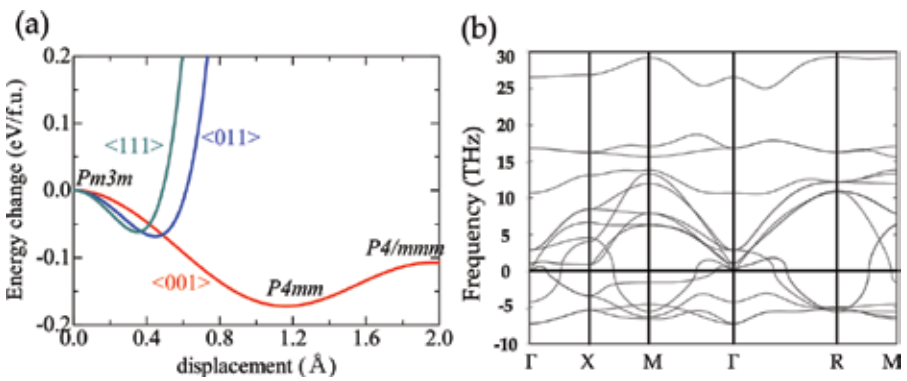

Figure 6. (a) Energy change of *Pm3m* phase as a function of li displacement along <001>, <011>, and <111> directions. (b) Phonon dispersion curve of *Pm3m* phase.

q-point	Frequency (THz)	Structure	Space group	Energy gain (eV/f.u.)
Γ	−7.22	Tetragonal	*P4mm*	−0.422
X	−5.38	Orthorhombic	*Pmma*	−0.220
M	−6.56	Orthorhombic	*Pmma*	−0.125
R	−5.51	Rhombohedral	*R-3c*	−0.682

"Structure" indicates Bravais lattice of modulated structure from *P4mm* phase. Space group of the relaxed modulated structure and energy gain is also shown.

Table 4. Summary of imaginary phonon frequency at each symmetric ***q***-point in ***P4mm*** phase of $LiNbO_3$ and corresponding structural phase transition.

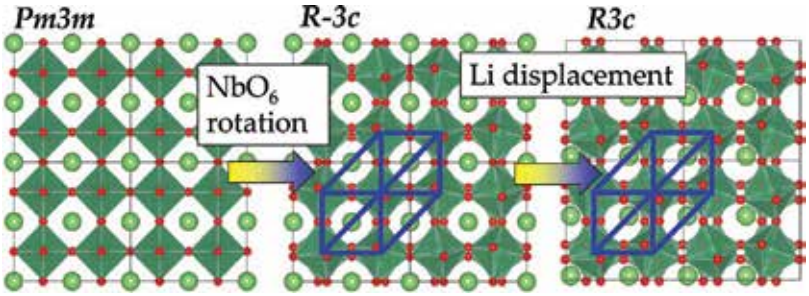

Figure 7. Phase transition path between cubic perovskite structure (*Pm3m*) and $LiNbO_3$-structure. Unit cell of $LiNbO_3$ is enclosed with blue lines.

Figure 7 shows schematic illustration of phase transition mechanism from cubic perovskite structure to $LiNbO_3$ structure. On the contrary to the result of **Figure 6a**, R_{25} soft mode is represented as rotation of NbO_6 polyhedra. Then, Γ_{15} soft mode of $R\bar{3}c$ phase leads *R3c* phase, which is ground state of $LiNbO_3$. Although the present study shows that perovskite-structured $LiNbO_3$ is thermodynamically unstable while its piezoelectricity is excellent, it can be possible to control phase transition behavior by dopant substitution.

7. Summary and conclusion

In this chapter, we briefly introduced sophisticated method of density functional perturbation theory. DFPT can effectively calculate the second derivative of the total energy with respect to the atomic displacement within the framework of first-principles calculation. By using DFPT method, we can predict piezoelectric constants, dielectric constants, elastic constants, and phonon dispersion relationship of any given crystal structure. Moreover, we showed our established computational technique to decompose piezoelectric constants into each atomic contribution, which enable us to gain deeper insights to understand the piezoelectricity of material. By using $LiNbO_3$ as a model material, we showed the predictability of DFPT for piezoelectric properties. In addition, we showed that superior piezoelectric properties are hidden in perovskite-structured $LiNbO_3$. Structural relationship and possible phase transition path between $LiNbO_3$ structure and perovskite structure were discussed and concluded that perovskite-structured $LiNbO_3$ is thermodynamically unstable. Further studies are expected to control relative phase stability between perovskite and $LiNbO_3$ structure by dopant substation and solid solution.

Author details

Kaoru Nakamura*, Sadao Higuchi and Toshiharu Ohnuma

*Address all correspondence to: n-kaoru@criepi.denken.or.jp

Central Research Institute of Electric Power Industry, Japan

References

[1] Martin Richard M. Piezoelectricity. Physical Review B. 1972;**5**:1607-1613. DOI: 10.1103/PhysRevB.5.1607

[2] Hohenberg P, Kohn W. Inhomogeneous electron gas. Physics Review. 1964;**136**:B864-B871. DOI: 10.1103/PhysRev.136.B864

[3] Kohn W, Sham LJ. Self-consistent equations including exchange and correlation effects. Physical Review. 1965;**140**:A1133-A1138. DOI: 10.1103/PhysRev.140.A1133

[4] King-Smith RD, Vanderbilt D. Theory of polarization of crystalline solids. Physical Review B. 1993;**47**:1651-1654. DOI: 10.1103/PhysRevB.47.1651

[5] Baroni S, Giannozzi P, Testa A. Green's-function approach to linear response in solids. Physical Review Letters. 1987;**58**:1861-1864. DOI: 10.1103/PhysRevLett.58.1861

[6] de Gironcoli S, Baroni S, Resta R. Piezoelectric properties of III-V semiconductors from first-principles linear-response theory. Physical Review Letters. 1989;**62**:2853-2856. DOI: 10.1103/PhysRevLett.62.2853

[7] Giannozzi P, de GS, Pavone P, Baroni S. Ab initio calculation of phonon dispersions in semiconductors. Physical Review B. 1991;**43**:7231-7242. DOI: 10.1103/PhysRevB.43.7231

[8] Gonze X. Perturbation expansion of variational principles at arbitrary order. Physical Review A. 1995;**52**:1096-1095. DOI: 10.1103/PhysRevA.52.1086

[9] Gonze X. First-principles responses of solids to atomic displacements and homogeneous electric fields: Implementation of a conjugate-gradient algorithm. Physical Review B. 1997;**55**:10337-10354. DOI: 10.1103/PhysRevB.55.10337

[10] Gonze X, Lee C. Dynamical matrices, born effective charges, dielectric permittivity tensors, and interatomic force constants from density-functional perturbation theory. Physical Review B. 1997;**55**:10355-10368. DOI: 10.1103/PhysRevB.55.10355

[11] Baroni S, de GS, Corso AD, Giannozzi P. Phonons and related crystal properties from density-functional perturbation theory. Reviews of Modern Physics. 2001;**73**:515-562. DOI: 10.1103/RevModPhys.73.515

[12] Sághi-Szabó G, Cohen RE. First-principles study of piezoelectricity in $PbTiO_3$. Physical Review Letters. 1998;**80**:4321-4324. DOI: 10.1103/PhysRevLett.80.4321

[13] Sághi-Szabó G, Cohen RE, Krakauer H. First-principles study of piezoelectricity in tetragonal $PbTiO_3$ and $PbZr_{1/2}Ti_{1/2}O_3$. Physical Review B. 1999;**59**:12771-12776. DOI: 10.1103/PhysRevB.59.12771

[14] Ghosez P, Cockayne E, Waghmare UV, Rabe KM. Lattice dynamics of $BaTiO_3$, $PbTiO_3$, and $PbZrO_3$: A comparative first-principles study. Physical Review B. 1999;**60**:836-843. DOI: 10.1103/PhysRevB.60.836

[15] Bellaiche L, Vanderbilt D. Intrinsic piezoelectric response in perovskite alloys: PMN-PT versus PZT. Physical Review Letters. 1999;**83**:1347-1350. DOI: 10.1103/PhysRevLett.83.1347

[16] Bernardini F, Fiorentini V, Vanderbilt D. Spontaneous polarization and piezoelectric constants of III-V nitrides. Physical Review B. 1997;**56**:R10024-R10027. DOI: 10.1103/PhysRevB.56.R10024

[17] Boonchun A, Lambrecht WRL. First-principles study of the elasticity, piezoelectricity, and vibrational styles in $LiGaO_2$ compared with ZnO and GaN. Physical Review B. 2010;**81**:235214. DOI: 10.1103/PhysRevB.81.235214

[18] Karanth D, Fu H. Large electromechanical response in ZnO and its microscopic origin. Physical Review B. 2005;**72**:064116. DOI: 10.1103/PhysRevB.72.064116

[19] Tasnádi F, Alling B, Höglund C, Wingqvist G, Birch J, Hultman L, Abrikosov IA. Origin of the anomalous piezoelectric response in Wurtzite $Sc_xAl_{1-x}N$ alloys. Physical Review Letters. 2010;**104**:137601. DOI: 10.1103/PhysRevLett.104.137601

[20] Al-Yacoub A, Bellaiche L, Wei SH. Piezoelectric coefficients of complex semiconductor alloys from first-principles: The case of $Ga_{1-x}In_xN$. Physical Review Letters. 2002;**89**: 057601. DOI: 10.1103/PhysRevLett.89.057601

[21] Armiento R, Kozinsky B, Fornari M, Ceder G. Screening for high-performance piezoelectrics using high-throughput density functional theory. Physical Review B. 2011;**84**: 014103. DOI: 10.1103/PhysRevB.84.014103

[22] Akgencab B, Kinacic A, Tassevend C, Cagin T. First-principles calculations on stability and mechanical properties of various ABO_3 and their alloys. Materials Chemistry and Physics. 2018;**205**:315-324. DOI: 10.1016/j.matchemphys.2017.11.026

[23] Dai JQ, Song YM, Zhang H. First-principles study of the phonon, dielectric, and piezoelectric response in Bi_2ZnTiO_6 supercell. Computational Materials Science. 2015;**101**:227-232. DOI: 10.1016/j.commatsci.2015.01.040

[24] Wan LF, Nishimatsu N, Beckman SP. The structural, dielectric, elastic, and piezoelectric properties of $KNbO_3$ from first-principles methods. Journal of Applied Physics. 2012;**111**:104107. DOI: 10.1063/1.4712052

[25] de Jong M, Chen W, Geerlings H, Asta M, Persson KA. A database to enable discovery and design of piezoelectric materials. Scientific Data. 2015;**2**:150053. DOI: 10.1038/sdata.2015.53

[26] Tinte S, Rabe KM, Vanderbilt D. Anomalous enhancement of tetragonality in $PbTiO_3$ induced by negative pressure. Physical Review B. 2003;**68**:144105. DOI: 10.1103/PhysRevB.68.144105

[27] Wu Z, Cohen RE. Pressure-induced anomalous phase transitions and colossal enhancement of piezoelectricity in $PbTiO_3$. Physical Review Letters. 2005;**95**:037601. DOI: 10.1103/PhysRevLett.95.037601

[28] Shahmirzaee H, Mardani R. Enhancement of piezoelectricity of tetragonal P4mm $SrHfO_3$ under uniaxial stress: A first principle study. Computational Condensed Matter. 2018;**14**:46-48. DOI: 10.1016/j.cocom.2017.12.006

[29] Duan Y, Lv D, Liu K, Wu H, Qin L, Shi L, Tang G. Strain-induced structural, band-structure and piezoelectric evolutions in $Al_{0.5}Ga_{0.5}N$ alloy. Journal of Applied Physics. 2015;**117**:045711. DOI: 10.1063/1.4906779

[30] Nakamura K, Higuchi S, Ohnuma T. Enhancement of piezoelectric constants induced by cation-substitution and two-dimensional strain effects on ZnO predicted by density functional perturbation theory. Journal of Applied Physics. 2016;**119**:114102. DOI: 10.1063/1.4943937

[31] Hamann DR, Wu X, Rabe KM, Vanderbilt D. Metric tensor formulation of strain in density-functional perturbation theory. Physical Review B. 2005;**71**:035117. DOI: 10.1103/PhysRevB.71.035117

[32] Wu X, Vanderbilt D, Hamann DR. Systematic treatment of displacements, strains, and electric fields in density-functional perturbation theory. Physical Review B. 2005;**72**:035105. DOI: 10.1103/PhysRevB.72.035105

[33] Gonze X, Jollet F, Araujo FA, Adams D, Amadon B, Applencourt T, Audouze C, Beuken J-M, Bieder J, Bokhanchuk A, Bousquet E, Bruneval F, Caliste D, Côté M, Dahm F, Da Pieve F, Delaveau M, Di Gennaro M, Dorado B, Espejo C, Geneste G, Genovese L, Gerossier A, Giantomassi M, Gillet Y, Hamann DR, He L, Jomard G, Janssen JL, Le RS, Levitt A, Lherbier A, Liu F, Lukacevic I, Martin A, Martins C, Oliveira MJT, Poncé S, Pouillon Y, Rangel T, Rignanese G-M, Romero AH, Rousseau B, Rubel O, Shukri AA, Stankovski M, Torrent M, Van Setten MJ, Van Troeye B, Verstraete MJ, Waroquier D, Wiktor J, Xue B, Zhou A, Zwanziger JW. Recent developments in the ABINIT software package. Computer Physics Communications. 2016;**205**:106-131. DOI: 10.1016/j.cpc.2016.04.003

[34] Kresse G, Furthmüller J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. Physical Review B. 1996;**54**:11169-11186. DOI: 10.1103/PhysRevB.54.11169

[35] Momma K, Izumi F. VESTA: A three-dimensional visualization system for electronic and structural analysis. Journal of Applied Crystallography. 2008;**41**:653-658. DOI: 10.1107/S0021889808012016

[36] Smolenskii GA, Krainik NN, Khuchua NP, Zhdanova VV, Mylnikova IE. The curie temperature of $LiNbO_3$. Physica Status Solidi B. 1966;**13**:309-314. DOI: 10.1002/pssb.19660130202

[37] Karapetyan KG, Kteyan AA, Vardanyan RA. Thermal reduction effect on curie temperature of $LiNbO_3$ ferroelectrics. Solid State Communications. 2006;**140**:474-476. DOI: 10.1016/j.ssc.2006.08.045

[38] Inaguma Y, Yoshida M, Tsuchiya T, Aimi A, Tanaka K, Katsumata T, Mori D. High-pressure synthesis of novel lithium niobate-type oxides. Journal of Physics Conference Series. 2010;**215**:012131. DOI: 10.1088/1742-6596/215/1/012131

[39] Inaguma Y, Yoshida M, Katsumata T. A polar oxide $ZnSnO_3$ with a $LiNbO_3$-type structure. Journal of the American Chemical Society. 2008;**130**:6704-6705. DOI: 10.1021/ja801843v

[40] Nakayama M, Nogami M, Yoshida M, Katsumata T, Inaguma Y. First-principles studies on novel polar oxide $ZnSnO_3$; pressure-induced phase transition and electric properties. Advanced Materials. 2010;**22**:2579-2582. DOI: 10.1002/adma.200903432

[41] Kuno Y, Tassel C, Fujita K, Batuk D, Abakumov AM, Shitara K, Kuwabara A, Moriwake H, Watabe D, Ritter C, Brown CM, Yamamoto T, Takeiri F, Abe R, Kobayashi Y, Tanaka K, Kageyama H. $ZnTaO_2N$: Stabilized high-temperature $LiNbO_3$-type structure. Journal of the American Chemical Society. 2016;**138**:15950-15955. DOI: 10.1021/jacs.6b08635

[42] Katsumata T, Ohba C, Tobe A, Takeda A, Shoji M, Aimi A, Mori D, Inaguma Y. Synthesis of new $LiNbO_3$-type oxynitrides, $Mn(Mn_{1/6}Ta_{5/6})O_{2.5}N_{0.5}$ under high pressure and at high temperature. Chemistry Letters. 2018;**47**:37-39. DOI: 10.1246/cl.170851

[43] Nakamura K, Higuchi S, Ohnuma T. First-principles investigation of pressure-induced phase transition in $LiNbO_3$. Journal of Applied Physics. 2012;**111**:033522. DOI: 10.1063/1.3682522

[44] Mukaide T, Yagi T, Miyajima N, Kondo T, Sata N. High pressure and high temperature phase transformations in $LiNbO_3$. Journal of Applied Physics. 2003;**93**:3582-3858. DOI: 10.1063/1.1556570

[45] Blöch PE. Projector augmented-wave method. Physical Review B. 1994;**50**:17953-17979. DOI: 10.1103/PhysRevB.50.17953

[46] Perdew JP, Ruzsinszky A, Csonka GI, Vydrov OA, Scuseria GE, Constantin LA, Zhou X, Burke K. Restoring the density-gradient expansion for exchange in solids and surfaces. Physics Review Letters. 2008;**100**:136406. DOI: 10.1103/PhysRevLett.100.136406

[47] Togo A, Oba F, Tanaka I. First-principles calculations of the ferroelastic transition between rutile-type and $CaCl_2$-type SiO_2 at high pressures. Physics Review B. 2008;**78**:134106. DOI: 10.1103/PhysRevB.78.134106

[48] Kushibiki J, Takanaga I, Arakawa M, Sannomiya T. Accurate measurements of the acoustical physical constants of $LiNbO_3$ and $LiTaO_3$ single crystals. IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control. 1999;**46**:1315-1323. DOI: 10.1109/58.796136

[49] Warner AW, Onoe M, Coquin GA. Determination of elastic and piezoelectric constants for crystals in class (3m). The Journal of the Acoustical Society of America. 1967;**42**:1223-1231. DOI: 10.1121/1.1910709

[50] Ghosez P, Michenaud JP, Gonze X. Dynamical atomic charges: The case of ABO_3 compounds. Physical Review B. 1998;**58**:6224-6240. DOI: 10.1103/PhysRevB.58.6224

[51] Kuwata J, Uchino K, Numura S. Dielectric and piezoelectric properties of $0.91Pb(Zn_{1/3}Nb_{2/3})O_3$-$0.09PbTiO_3$ single crystals. Japanese Journal of Applied Physics. 1982;**21**:1298-1301. DOI: 10.1143/JJAP.21.1298

[52] Guo Y, Kakimoto K, Ohsato H. Phase transitional behavior and piezoelectric properties of $(Na_{0.5}K_{0.5})NbO_3$–$LiNbO_3$ ceramics. Applied Physics Letters. 2004;**85**:4121-4123. DOI: 10.1063/1.1813636

Chapter 2

Sliding-Mode Perturbation Observer-Based Sliding-Mode Control for VSC-HVDC Systems

Bo Yang, Tao Yu, Hongchun Shu and Pulin Cao

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.74717

Abstract

This chapter develops a sliding-mode perturbation observer-based sliding-mode control (POSMC) scheme for voltage source converter-based high voltage direct current (VSC-HVDC) systems. The combinatorial effect of nonlinearities, parameter uncertainties, unmodeled dynamics, and time-varying external disturbances is aggregated into a perturbation, which is estimated online by a sliding-mode state and perturbation observer (SMSPO). POSMC does not require an accurate VSC-HVDC system model and only the reactive power and DC voltage at the rectifier side while reactive and active powers at the inverter side need to be measured. Additionally, a considerable robustness can be provided through the real-time compensation of the perturbation, in which the upper bound of perturbation is replaced by the real-time estimation of the perturbation, such that the over-conservativeness of conventional sliding-mode control (SMC) can be effectively reduced. Four case studies are carried out on the VSC-HVDC system, such as active and reactive power tracking, AC bus fault, system parameter uncertainties, and weak AC gird connection. Simulation results verify its advantages over vector control and feedback linearization sliding-mode control. Then, a dSPACE-based hardware-in-the-loop (HIL) test is undertaken to validate the implementation feasibility of the proposed approach.

Keywords: sliding-mode control, sliding-mode perturbation observer, VSC-HVDC systems, HIL test

1. Introduction

In the past decades, the ever-increasing penetration of renewable energy (wind, solar, wave, hydro, and biomass) requires an extraordinarily reliable and effective transmission of electrical power from these new sources to the main power grid [1], in which hydropower has already

been fully exploited in many grids, such that a sustainable development can be achieved in future [2]. The problems and perspectives of converting present energy systems (mainly thermal and nuclear) into a 100% renewable energy system have been discussed with a conclusion that such idea is possible, which, however, raises that advanced transmission technologies are needed to realize this goal [3].

The need for more secure power grids and ever-increasing environmental concerns continue to drive the worldwide deployment of high voltage direct current (HVDC) transmission technology, which enables a more reliable and stable asynchronous interconnection of power networks that operate on different frequencies [4]. HVDC systems use power electronic devices to convert alternative current (AC) into direct current (DC), they are an economical way of transmitting bulk electrical power in DC over long distance overhead line or short submarine cable, while advanced extruded DC cable technologies have been used to increase power transmissions by at least 50%, which is also an important onshore solution. HVDC enables secure and stable asynchronous interconnection of power networks that operate on different frequencies. Different technologies have been used to design two-terminal HVDC systems for the purpose of a point-to-point power transfer, such as line-commutated converter (LCC)-based HVDC (LCC-HVDC) systems using grid-controlled mercury-arc valves or thyristors, capacitor-commutated converter (CCC)-based HVDC (CCC-HVDC) systems, or controlled series commutated converter (CSCC)-based HVDC (CSCC-HVDC) systems [5].

Voltage source converter-based high voltage direct current (VSC-HVDC) systems using insulated gate bipolar transistor (IGBT) technology have attracted increasing attentions due to the interconnection between the mainland and offshore wind farms, power flow regulation in alternating current (AC) power systems, long distance transmission [6], and introduction of the supergrid, which is a large-scale power grid interconnected between national power grids [7]. The main feature of the VSC-HVDC system is that no external voltage source is needed for communication, while active and reactive powers at each AC grid can be independently controlled [8, 9].

Traditionally, control of the VSC-HVDC system utilizes a nested-loop d-q vector control (VC) approach based on linear proportional-integral (PI) methods [10], whose control performance may be degraded with the change of operation conditions as its control parameters are tuned from one-point linearization model [11]. As VSC-HVDC systems are highly nonlinear resulting from converters and also operate in power systems with modeling uncertainties, many advanced control approaches are developed to provide a consistent control performance under various operation conditions, such as feedback linearization control (FLC) [12], which fully compensated the nonlinearities with the requirement of an accurate system model. Linear matrix inequality (LMI)-based robust control was developed in [13] to maximize the size of the uncertainty region within which closed-loop stability is maintained. In addition, adaptive backstepping control was designed to estimate the uncertain parameters by [14]. In [8, 9], power-synchronization control was employed to greatly increase the short-circuit capacity to the AC system. However, the aforementioned methods may not be adequate to simultaneously handle perturbations such as modeling uncertainties and time-varying external disturbances.

Based on the variable structure control strategy, sliding-mode control (SMC) is an effective and high-frequency switching control for nonlinear systems with modeling uncertainties and time-varying external disturbances. The main idea of SMC is to maintain the system sliding on a surface

in the state space via an appropriate switching logic; it features the simple implementation, disturbance rejection, fast response, and strong robustness [15]. While the malignant effect of chattering phenomenon can be reduced by predictive variable structure [16] and self-tuning sliding mode [17], SMC has been applied on electrical vehicles [18], power converters [19], induction machines [20], wind turbines [21], etc. Moreover, a feedback linearization sliding-mode control (FLSMC) has been developed for the VSC-HVDC system to offer invariant stability to modeling uncertainties by [22]. Basically, SMC assumes perturbations to be bounded and the prior knowledge of these upper bounds is required. However, it may be difficult or sometimes impossible to obtain these upper bounds, thus the supreme upper bound is chosen to cover the whole range of perturbations. As a consequence, SMC based on this knowledge becomes over-conservative which may cause a poor tracking performance and undesirable control oscillations [23].

During the past decades, several elegant approaches based on observers have been proposed to estimate perturbations, including the unknown input observer (UIO) [24], the disturbance observer (DOB) [25], the equivalent input disturbance (EID)-based estimation [26], enhanced decentralized PI control via advanced disturbance observer [27], the extended state observer (ESO)-based active disturbance rejection control (ADRC) [28], and practical multivariable control based on inverted decoupling and decentralized ADRC [29]. Among the above listed approaches, ESO requires the least amount of system information, in fact, only the system order needs to be known [30]. Due to such promising features, ESO-based control schemes have become more and more popular. Recently, ESO-based SMC has been developed to remedy the over-conservativeness of SMC via an online perturbation estimation. It observes both system states and perturbations by defining an extended state to represent the lumped perturbation, which can be then compensated online to improve the performance of system. Related applications can be referred to mechanical systems [31], missile systems [32], spherical robots [33], and DC-DC buck power converters [34].

This chapter uses an ESO called sliding-mode state and perturbation observer (SMSPO) [35, 36] to estimate the combinatorial effect of nonlinearities, parameter uncertainties, unmodeled dynamics, and time-varying external disturbances existed in VSC-HVDC systems, which is then compensated by the perturbation observer-based sliding-mode control (POSMC). The motivation to use POSMC, in this chapter, rather than SMC and our previous work [35–37] can be summarized as follows:

- The robustness of POSMC to the perturbation mostly depends on the perturbation compensation, while the ground of the robustness in SMC [18–22] is the discrete switching input. Furthermore, the upper bound of perturbation is replaced by the smaller bound of its estimation error, thus an over-conservative control input is avoided and the tracking accuracy is improved.
- POSMC can provide greater robustness than that of nonlinear adaptive control (NAC) [35, 36] and perturbation observer-based adaptive passive control (POAPC) [37] due to its inherent property of disturbance rejection.

Compared to VC [11], POSMC can provide a consistent control performance under various operation condition of the VSC-HVDC system and improve the power tracking by eliminating the power overshoot. Compared to FLSMC [22], POSMC only requires the measurement of

active and reactive power and DC voltage, which can provide a significant robustness and avoid an over-conservative control input as the real perturbation is estimated and compensated online. Four case studies are carried out to evaluate the control performance of POSMC through simulation, such as active and reactive power tracking, AC bus fault, system parameter uncertainties, and weak AC gird connection. Compared to the author's previous work on SMSPO [35, 36], a dSPACE simulator-based hardware-in-the-loop (HIL) test is undertaken to validate its implementation feasibility.

The rest of the chapter is organized as follows. In Section 2, the model of the two-terminal VSC-HVDC system is presented. In Section 3, POSMC design for the VSC-HVDC system is developed and discussed. Sections 4 and 5 present the simulation and HIL results, respectively. Finally, conclusions are drawn in Section 6.

2. VSC-HVDC system modeling

There are two VSCs in the VSC-HVDC system shown in **Figure 1**, in which the rectifier regulates the DC voltage and reactive power, while the inverter regulates the active and reactive power. Only the balanced condition is considered, e.g., the three phases have identical parameters and their voltages and currents have the same amplitude while each phase shifts 120° between themselves. The rectifier dynamics can be written at the angular frequency ω as [14].

$$\begin{cases} \dfrac{\mathrm{d}i_{\mathrm{d1}}}{\mathrm{d}t} = -\dfrac{R_1}{L_1} i_{\mathrm{d1}} + \omega i_{\mathrm{q1}} + u_{\mathrm{d1}} \\ \dfrac{\mathrm{d}i_{\mathrm{q1}}}{\mathrm{d}t} = -\dfrac{R_1}{L_1} i_{\mathrm{q1}} - \omega i_{\mathrm{d1}} + u_{\mathrm{q1}} \\ \dfrac{\mathrm{d}V_{\mathrm{dc1}}}{\mathrm{d}t} = \dfrac{3u_{\mathrm{sq1}} i_{\mathrm{q1}}}{2C_1 V_{\mathrm{dc1}}} - \dfrac{i_{\mathrm{L}}}{C_1} \end{cases} \tag{1}$$

where the rectifier is connected with the AC grid via the equivalent resistance and inductance R_1 and L_1, respectively. C_1 is the DC bus capacitor, $u_{\mathrm{d1}} = \frac{u_{\mathrm{sd1}} - u_{\mathrm{rd}}}{L_1}$ and $u_{\mathrm{q1}} = \frac{u_{\mathrm{sq1}} - u_{\mathrm{rq}}}{L_1}$.

The inverter dynamics is written as

$$\begin{cases} \dfrac{\mathrm{d}i_{\mathrm{d2}}}{\mathrm{d}t} = -\dfrac{R_2}{L_2} i_{\mathrm{d2}} + \omega i_{\mathrm{q2}} + u_{\mathrm{d2}} \\ \dfrac{\mathrm{d}i_{\mathrm{q2}}}{\mathrm{d}t} = -\dfrac{R_2}{L_2} i_{\mathrm{q2}} - \omega i_{\mathrm{d2}} + u_{\mathrm{q2}} \\ \dfrac{\mathrm{d}V_{\mathrm{dc2}}}{\mathrm{d}t} = \dfrac{3u_{\mathrm{sq2}} i_{\mathrm{q2}}}{2C_2 V_{\mathrm{dc2}}} + \dfrac{i_{\mathrm{L}}}{C_2} \end{cases} \tag{2}$$

where the inverter is connected with the AC grid via the equivalent resistance and inductance R_2 and L_2, respectively. C_2 is the DC bus capacitor, $u_{\mathrm{d2}} = \frac{u_{\mathrm{sd2}} - u_{\mathrm{id}}}{L_2}$ and $u_{\mathrm{q2}} = \frac{u_{\mathrm{sq2}} - u_{\mathrm{iq}}}{L_2}$.

The interconnection between the rectifier and inverter through DC cable is given as

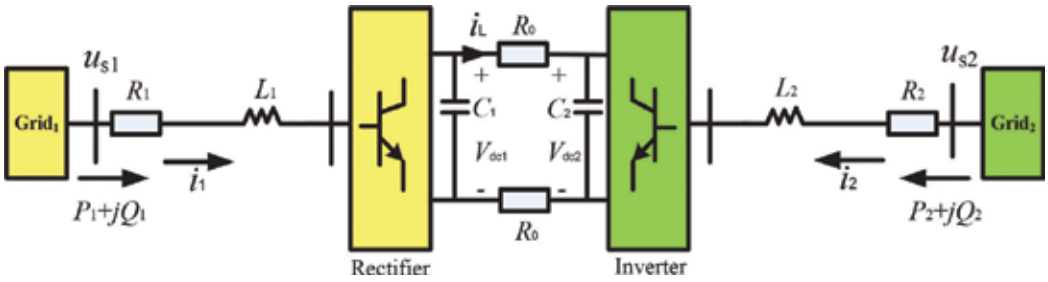

Figure 1. A standard two-terminal VSC-HVDC system.

$$V_{dc1} i_L = V_{dc2} i_L + 2R_0 i_L^2 \tag{3}$$

where R_0 represents the equivalent DC cable resistance.

The phase-locked loop (PLL) [38] is used during the transformation of the *abc* frame to the *dq* frame. In the synchronous frame, u_{sd1}, u_{sd2}, u_{sq1}, and u_{sq2} are the *d*, *q* axes components of the respective AC grid voltages; i_{d1}, i_{d2}, i_{q1}, and i_{q2} are that of the line currents; u_{rd}, u_{id}, u_{rq}, and u_{iq} are that of the converter input voltages. P_1, P_2, Q_1, and Q_2 are the active and reactive powers transmitted from the AC grid to the VSC; V_{dc1} and V_{dc2} are the DC voltages; and i_L is the DC cable current.

At the rectifier side, the *q*-axis is set to be in phase with the AC grid voltage u_{s1}. Correspondingly, the *q*-axis is set to be in phase of the AC grid voltage u_{s2} at the inverter side. Hence, u_{sd1} and u_{sd2} are equal to 0, while u_{sq1} and u_{sq2} are equal to the magnitude of u_{s1} and u_{s2}. Note that this chapter adopts such framework from [12, 14, 22] to provide a consistent control design procedure and an easy control performance comparison, other framework can also be used as shown in [8, 9, 11]. The only difference of these two alternatives is the derived system equations, while the control design is totally the same. In addition, it is assumed that the VSC-HVDC system is connected to sufficiently strong AC grids, such that the AC grid voltage remains as an ideal constant. The power flows from the AC grid can be given as

$$\begin{cases} P_1 = \dfrac{3}{2}\left(u_{sq1} i_{q1} + u_{sd1} i_{d1}\right) = \dfrac{3}{2} u_{sq1} i_{q1} \\ Q_1 = \dfrac{3}{2}\left(u_{sq1} i_{d1} - u_{sd1} i_{q1}\right) = \dfrac{3}{2} u_{sq1} i_{d1} \\ P_2 = \dfrac{3}{2}\left(u_{sq2} i_{q2} + u_{sd2} i_{d2}\right) = \dfrac{3}{2} u_{sq2} i_{q2} \\ Q_2 = \dfrac{3}{2}\left(u_{sq2} i_{d2} - u_{sd2} i_{q2}\right) = \dfrac{3}{2} u_{sq2} i_{d2} \end{cases} \tag{4}$$

3. POSMC design for the VSC-HVDC system

3.1. Perturbation observer-based sliding-mode control

Consider an uncertain nonlinear system which has the following canonical form:

$$\begin{cases} \dot{x} = Ax + B(a(x) + b(x)u + d(t)) \\ y = x_1 \end{cases} \tag{5}$$

where $x = [x_1, x_2, \cdots, x_n]^{\mathrm{T}} \in \mathbb{R}^n$ is the state variable vector, $u \in \mathbb{R}$ and $y \in \mathbb{R}$ are the control input and system output, respectively. $a(x)$: $\mathbb{R}^n \mapsto \mathbb{R}$ and $b(x)$: $\mathbb{R}^n \mapsto \mathbb{R}$ are unknown smooth functions, and $d(t)$: $\mathbb{R}^+ \mapsto \mathbb{R}$ represents the time-varying external disturbance. The $n \times n$ matrix A and the $n \times 1$ matrix B are of the canonical form as follows:

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}_{n \times n}, \qquad B = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}_{n \times 1} \tag{6}$$

The perturbation of system (5) is defined as [35–37]

$$\Psi(x, u, t) = a(x) + (b(x) - b_0)u + d(t) \tag{7}$$

From the original system (5), the last state x_n can be rewritten in the presence of perturbation (6) as follows:

$$\dot{x}_n = a(x) + (b(x) - b_0)u + d(t) + b_0 u = \Psi(x, u, t) + b_0 u \tag{8}$$

Define a *fictitious state* $x_{n+1} = \Psi(x, u, t)$. Then, system (5) can be extended as

$$\begin{cases} y = x_1 \\ \dot{x}_1 = x_2 \\ \vdots \\ \dot{x}_n = x_{n+1} + b_0 u \\ \dot{x}_{n+1} = \dot{\Psi}(\cdot) \end{cases} \tag{9}$$

The new state vector becomes $x_{\mathrm{e}} = [x_1, x_2, \cdots, x_n, x_{n+1}]^{\mathrm{T}}$, and following assumptions are made [35]:

- **A.1** b_0 is chosen to satisfy: $|b(x)/b_0 - 1| \leq \theta < 1$, where θ is a positive constant.
- **A.2** The functions $\Psi(x, u, t) : \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^+ \mapsto \mathbb{R}$ and $\dot{\Psi}(x, u, t) : \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^+ \mapsto \mathbb{R}$ are bounded over the domain of interest: $|\Psi(x, u, t)| \leq \gamma_1$, $|\dot{\Psi}(x, u, t)| \leq \gamma_2$ with $\Psi(0, 0, 0) = 0$ and $\dot{\Psi}(0, 0, 0) = 0$, where γ_1 and γ_2 are positive constants.
- **A.3** The desired trajectory y_{d} and its up to nth-order derivative are continuous and bounded.

The above three assumptions ensure the effectiveness of such perturbation estimation-based approach. In particular, assumptions A.1 and A.2 guarantee the closed-loop system

stability with perturbation estimation, while assumption A.3 ensures POSMC can drive the system state x to track a desired state $x_d = \left[y_d, y_d^{(1)}, \cdots, y_d^{(n-1)}\right]^T$ [39]. In the consideration of the worst case, e.g., $y = x_1$ is the only measurable state, an (n+1)th-order SMSPO [35, 36] for the extended system (8) is designed to estimate the system states and perturbation, shown as follows:

$$\begin{cases} \dot{\hat{x}}_1 = \hat{x}_2 + \alpha_1 \tilde{x}_1 + k_1 \text{sat}(\tilde{x}_1) \\ \quad \vdots \\ \dot{\hat{x}}_n = \hat{\Psi}(\cdot) + \alpha_n \tilde{x}_1 + k_n \text{sat}(\tilde{x}_1) + b_0 u \\ \dot{\hat{\Psi}}(\cdot) = \alpha_{n+1} \tilde{x}_1 + k_{n+1} \text{sat}(\tilde{x}_1) \end{cases} \tag{10}$$

where $\tilde{x}_1 = x_1 - \hat{x}_1$, k_i and α_i, $i = 1, 2, \cdots, n+1$, are positive coefficients, function $\text{sat}(\tilde{x}_1)$ is defined as $\text{sat}(\tilde{x}_1) = \tilde{x}_1/|\tilde{x}_1|$ when $|\tilde{x}_1| > \epsilon$ and $\text{sat}(\tilde{x}_1) = \tilde{x}_1/\epsilon$ when $|\tilde{x}_1| \leq \epsilon$. The effect and setting of the SMSPO parameters are provided as follows:

- **The Luenberger observer constants** α_i, which are chosen to place the observer poles at the desired locations in the open left-half complex plane. In other words, α_i are chosen such that the root of $s^{n+1} + \alpha_1 s^n + \alpha_2 s^{n-1} + \cdots + \alpha_{n+1} = (s + \lambda_\alpha)^{n+1} = 0$ is in the open left-half complex plane. A larger value of α_i not only will accelerate the estimation rate of SMSPO, but also will result in a more significant effect of peaking phenomenon. Thus, a trade-off between the estimation rate and effect of peaking phenomenon must be made through trial-and-error. Normally, they are set to be much larger than the root of the closed-loop system to ensure a fast online estimation [37].

- **The sliding surface constants** k_i. $k_1 \geq |\tilde{x}_2|_{\max}$ must be chosen to guarantee the estimation error of SMSPO (9) will enter into the sliding surface $S_{spo}(\tilde{x}) = \tilde{x}_1 = 0$ at $t > t_s$ and thereafter remain $S_{spo} = 0$, $t \geq t_s$ [35, 39]. While the poles of the sliding surface λ_k are determined by choosing the ratio $k_i/k_1 (i = 2, 3, \cdots, n+1)$ to put the root of $p^n + (k_2/k_1)p^{n-1} + \cdots + (k_n/k_1)p + (k_{n+1}/k_1) = (p + \lambda_k)^n = 0$ to be in the open left-half complex plane. Under Assumption A.2, SMSPO converges to a neighborhood of the origin if gains k_i are properly selected, which has been proved in [35, 40]. For a given k_1, a larger k_i not only will accelerate the estimation rate of SMSPO, but also will result in a degraded observer stability. Thus, a trade-off between the estimation rate and observer stability must be made through trial-and-error [39].

- **The layer thickness constant of saturation function** ϵ, which is a positive small scaler to replace the sign function by the saturation function, such that the chattering effect can be reduced. A larger ϵ will result in a smoother chattering, but a larger steady-state estimation error. Consequently, a trade-off between the chattering effect and steady-state estimation error must be made through trial-and-error. In practice, a value closes to 0 is recommended.

Moreover, the reduced estimation error dynamics on the sliding mode can be written as [35]

$$\begin{cases} \dot{\tilde{x}}_2 = -\dfrac{k_2}{k_1}\tilde{x}_2 + \tilde{x}_3 \\ \dot{\tilde{x}}_3 = -\dfrac{k_3}{k_1}\tilde{x}_2 + \tilde{x}_4 \\ \quad\vdots \\ \dot{\tilde{x}}_n = -\dfrac{k_n}{k_1}\tilde{x}_2 + \tilde{\Psi}(\cdot) \\ \dot{\tilde{\Psi}}(\cdot) = -\dfrac{k_{n+1}}{k_1}\tilde{x}_2 + \dot{\Psi}(\cdot) \end{cases} \tag{11}$$

Lemma 1 [39]. Consider extended system (8), design an SMSPO (9). If assumption A.2 holds for some value γ_2, then given any constant δ, the gains k_i can be chosen such that, from an initial estimation error $\tilde{x}_{\mathrm{e}}(0)$, the estimation error $\tilde{x}_{\mathrm{e}}$ converges exponentially into the neighborhood

$$\|\tilde{x}_{\mathrm{e}}\| \leq \delta \tag{12}$$

In particular,

$$|\tilde{x}_i| \leq \frac{\delta}{\lambda_{\mathrm{k}}^{n+1-i}}, \quad i = 2, \cdots, n+1, \quad \forall t > t_1. \tag{13}$$

where t_1 is the time constant which definition can be found in [39].

Remark 1. When SMSPO is used to estimate the perturbation, the upper bound of the derivative of perturbation γ_2 is required to guarantee the estimation accuracy, and such upper bound will result in a conservative observer gain. However, the conservative gain is only included in the observer loop, not in the controller loop.

Define an estimated sliding surface as

$$\widehat{S}(x,t) = \sum_{i=1}^{n} \rho_i \left(\widehat{x}_i - y_{\mathrm{d}}^{(i-1)} \right) \tag{14}$$

where the estimated sliding surface gains $\rho_i = C_{n-1}^{i-1} \lambda_{\mathrm{c}}^{n-i}$, $i = 1, \cdots, n$, place all poles of the estimated sliding surface at $-\lambda_{\mathrm{c}}$, where $\lambda_{\mathrm{c}} > 0$.

Besides, the actual sliding surface is written by

$$S = \sum_{i=1}^{n} \rho_i \left(x_i - y_{\mathrm{d}}^{(i-1)} \right) \tag{15}$$

Hence, the estimation error of the sliding surface can be directly calculated as

$$\tilde{S} = S - \widehat{S} = \sum_{i=1}^{n} \rho_i \tilde{x}_i \tag{16}$$

Construct a Lyapunov function as follows:

$$V = \frac{1}{2}\widehat{S}^2 \tag{17}$$

The attractiveness of sliding surface is achieved if $\dot{V} < 0$ for all $\tilde{x} \notin \widehat{S}$, that is, the control u needs to be designed to enforce $\widehat{S}\dot{\widehat{S}} < 0$ outside a prescribed manifold $|\widehat{S}| < \varepsilon_c$.

The POSMC for system (5) is designed as

$$u = \frac{1}{b_0}\left[y_d^{(n)} - \sum_{i=1}^{n-1}\rho_i\left(\widehat{x}_{i+1} - y_d^{(i)}\right) - \zeta\widehat{S} - \varphi \mathrm{sat}\left(\widehat{S}\right) - \widehat{\Psi}(\cdot)\right] \tag{18}$$

where ζ and φ are control gains which are chosen to fulfill the attractiveness of the estimated sliding surface $\widehat{S}$.

Differentiate estimated sliding surface (13) along SMSPO (9), use the reduced estimation error dynamics (10), it yields

$$\dot{\widehat{S}} = \widehat{\Psi}(\cdot) + b_0 u + \frac{k_n}{k_1}\tilde{x}_2 - y_d^{(n)} + \sum_{i=1}^{n-1}\rho_i\left(\widehat{x}_{i+1} - y_d^{(i)} + \frac{k_i}{k_1}\tilde{x}_2\right) \tag{19}$$

Substitute control (17) into the above Eq. (18), leads to

$$\dot{\widehat{S}} = \sum_{i=1}^{n}\rho_i\frac{k_i}{k_1}\tilde{x}_2 - \zeta\widehat{S} - \varphi \mathrm{sat}\left(\widehat{S}, \varepsilon_c\right) \tag{20}$$

Consequently, the attractiveness of sliding surface can be derived as

$$\zeta|\widehat{S}| + \varphi > \sum_{i=1}^{n}\rho_i\frac{k_i}{k_1}|\tilde{x}_2| \tag{21}$$

which will be fulfilled with the relationship of k_1 if

$$\zeta|\widehat{S}| + \varphi > k_1\sum_{i=1}^{n}\rho_i\frac{k_i}{k_1} \tag{22}$$

The above condition can be immediately satisfied if control gain φ is chosen as

$$\varphi > k_1\sum_{i=1}^{n}\rho_i\frac{k_i}{k_1} \tag{23}$$

which, using gains k_i, yields

$$\varphi > k_1\sum_{i=1}^{n}\rho_i C_n^{i-1}\lambda_k^{i-1} \tag{24}$$

This condition ensures the existence of a sliding mode on the boundary layer $|\widehat{S}| \leq \varepsilon_c$. From system (15) one can easily calculate

$$\dot{\tilde{S}} = \sum_{i=1}^{n-1} \rho_i \tilde{x}_{i+1} - \sum_{i=1}^{n} \rho_i \frac{k_i}{k_1} \tilde{x}_2 + \tilde{\Psi}(\cdot) \tag{25}$$

As $\widehat{S} = S - \tilde{S}$, the actual S-dynamics of sliding surface can be obtained with dynamics (19) as

$$\dot{S} + \left(\zeta + \frac{\varphi}{\varepsilon_c}\right) S = \left(\zeta + \frac{\varphi}{\varepsilon_c}\right) \sum_{i=1}^{n} \rho_i \tilde{x}_i + \sum_{i=1}^{n-1} \rho_i \tilde{x}_{i+1} + \tilde{\Psi}(\cdot) \tag{26}$$

It is definite that the driving term of S-dynamics is the sum of the estimation errors of states and the perturbation. The bounds of the sliding surface can be calculated by

$$|\widehat{S}| \leq \varepsilon_c \Rightarrow |S - \tilde{S}| \leq \varepsilon_c \Rightarrow |S| \leq |\tilde{S}| + \varepsilon_c \Rightarrow |S| \leq |\sum_{i=1}^{n} \rho_i \tilde{x}_i| + \varepsilon_c \leq \frac{\delta}{\lambda_k^{n+1}} \sum_{i=2}^{n} \rho_i \lambda_k^i + \varepsilon_c, \quad \forall t > t_1. \tag{27}$$

Based on bounds (26), together with the polynomial gains ρ_i, the states tracking error satisfies the following relationship [41]

$$|x^{(i)}(t) - x_d^{(i)}(t)| \leq (2\lambda_c)^i \frac{\varepsilon_c}{\lambda_c^n} + \frac{\delta}{\lambda_k^{n+1}} \sum_{j=2}^{n} \left(\frac{\lambda_k}{\lambda_c}\right)^j C_{n-1}^j, \qquad i = 0, 1, \cdots, n-1. \tag{28}$$

Note that POSMC does not require an accurate system model, and only one state measurement $y = x_1$ is needed. As the upper bound of perturbation $\Psi(\cdot)$ is replaced by the smaller bound of its estimation error $\tilde{\Psi}(\cdot)$, a smaller control gain is needed such that the over-conservativeness of SMC can be avoided [35].

Remark 2. The motivation to use SMSPO is due to the fact that the sliding-mode observer potentially offers advantages similar to those of sliding-mode controllers, in particular, inherent robustness to parameter uncertainty and external disturbances [42]. It is a high-performance state estimator with a simple structure and is well suited for uncertain nonlinear systems [31]. Moreover, it has the merits of simple structure and easy analysis of the closed-loop system stability compared to that of ADRC which uses a nonlinear observer [28], while they can provide almost the same performance of perturbation estimation.

The overall design procedure of POSMC for system (5) can be summarized as follows:

Step 1. Define perturbation (6) for the original nth-order system (5);

Step 2. Define a *fictitious state* $x_{n+1} = \Psi(\cdot)$ to represent perturbation (6);

Step 3. Extend the original nth-order system (5) into the extended $(n+1)$th-order system (8);

Step 4. Design the $(n+1)$th-order SMSPO (9) for the extended $(n+1)$th-order system (8) to obtain the state estimate $\widehat{x}$ and the perturbation estimate $\widehat{\Psi}(\cdot)$ by the only measurement of x_1;

Step 5. Design controller (17) for the original nth-order system (5), in which the estimated sliding surface $\widehat{S}$ is calculated by (13).

3.2. Rectifier controller design

Choose the system output $y_r = [y_{r1}, y_{r2}]^T = [Q_1, V_{dc1}]^T$, let Q_1^* and V_{dc1}^* be the given references of the reactive power and DC voltage, respectively. Define the tracking error $e_r = [e_{r1}, e_{r2}]^T = [Q_1 - Q_1^*, V_{dc1} - V_{dc1}^*]^T$, differentiate e_r for rectifier (1) until the control input appears explicitly, yields

$$\begin{bmatrix} \dot{e}_{r1} \\ \ddot{e}_{r2} \end{bmatrix} = \begin{bmatrix} f_{r1} - \dot{Q}_1^* \\ f_{r2} - \ddot{V}_{dc1}^* \end{bmatrix} + B_r \begin{bmatrix} u_{d1} \\ u_{q1} \end{bmatrix} \tag{29}$$

where

$$\begin{cases} f_{r1} &= \dfrac{3u_{sq1}}{2}\left(-\dfrac{R_1}{L_1} i_{d1} + \omega i_{q1}\right) \\ f_{r2} &= \dfrac{3u_{sq1}}{2C_1 V_{dc1}}\left[-\omega i_{d1} - \dfrac{R_1}{L_1} i_{q1} - \dfrac{i_{q1}}{V_{dc1}}\left(\dfrac{3u_{sq1} i_{q1}}{2C_1 V_{dc1}} - \dfrac{i_L}{C_1}\right)\right] \\ & \quad - \dfrac{1}{2R_0 C_1}\left(\dfrac{3u_{sq1} i_{q1}}{2C_1 V_{dc1}} - \dfrac{i_L}{C_1} - \dfrac{3u_{sq2} i_{q2}}{2C_2 V_{dc2}} - \dfrac{i_L}{C_2}\right) \end{cases} \tag{30}$$

and

$$B_r = \begin{bmatrix} \dfrac{3u_{sq1}}{2L_1} & 0 \\ 0 & \dfrac{3u_{sq1}}{2C_1 L_1 V_{dc1}} \end{bmatrix} \tag{31}$$

The determinant of matrix B_r is obtained as $|B_r| = 9u_{sq1}^2/\left(4C_1 L_1^2 V_{dc1}\right)$, which is nonzero within the operation range of the rectifier, thus system (28) is linearizable.

Assume all the nonlinearities are unknown, define the perturbations $\Psi_{r1}(\cdot)$ and $\Psi_{r2}(\cdot)$ as

$$\begin{bmatrix} \Psi_{r1}(\cdot) \\ \Psi_{r2}(\cdot) \end{bmatrix} = \begin{bmatrix} f_{r1} \\ f_{r2} \end{bmatrix} + (B_r - B_{r0}) \begin{bmatrix} u_{d1} \\ u_{q1} \end{bmatrix} \tag{32}$$

where the constant control gain B_{r0} is given by

$$B_{r0} = \begin{bmatrix} b_{r10} & 0 \\ 0 & b_{r20} \end{bmatrix} \tag{33}$$

Then system (28) can be rewritten as

$$\begin{bmatrix} \dot{e}_{r1} \\ \ddot{e}_{r2} \end{bmatrix} = \begin{bmatrix} \Psi_{r1}(\cdot) \\ \Psi_{r2}(\cdot) \end{bmatrix} + B_{r0} \begin{bmatrix} u_{d1} \\ u_{q1} \end{bmatrix} - \begin{bmatrix} \dot{Q}_1^* \\ \ddot{V}_{dc1}^* \end{bmatrix} \tag{34}$$

Define $z'_{11} = Q_1$, a second-order sliding-mode perturbation observer (SMPO) is used to estimate $\Psi_{r1}(\cdot)$ as

$$\begin{cases} \dot{\hat{z}}'_{11} = \widehat{\Psi}_{r1}(\cdot) + \alpha'_{r1}\tilde{Q}_1 + k'_{r1}\mathrm{sat}(\tilde{Q}_1) + b_{r10}u_{d1} \\ \dot{\widehat{\Psi}}_{r1}(\cdot) = \alpha'_{r2}\tilde{Q}_1 + k'_{r2}\mathrm{sat}(\tilde{Q}_1) \end{cases} \tag{35}$$

where observer gains k'_{r1}, k'_{r2}, α'_{r1}, and α'_{r2} are all positive constants.

Define $z_{11} = V_{dc1}$ and $z_{12} = \dot{z}_{11}$, a third-order SMSPO is used to estimate $\Psi_{r2}(\cdot)$ as

$$\begin{cases} \dot{\hat{z}}_{11} = \hat{z}_{12} + \alpha_{r1}\tilde{V}_{dc1} + k_{r1}\mathrm{sat}(\tilde{V}_{dc1}) \\ \dot{\hat{z}}_{12} = \widehat{\Psi}_{r2}(\cdot) + \alpha_{r2}\tilde{V}_{dc1} + k_{r2}\mathrm{sat}(\tilde{V}_{dc1}) + b_{r20}u_{q1} \\ \dot{\widehat{\Psi}}_{r2}(\cdot) = \alpha_{r3}\tilde{V}_{dc1} + k_{r3}\mathrm{sat}(\tilde{V}_{dc1}) \end{cases} \tag{36}$$

where observer gains k_{r1}, k_{r2}, k_{r3}, α_{r1}, α_{r2}, and α_{r3} are all positive constants.

The above observers (31) and (32) only need the measurement of reactive power Q_1 and DC voltage V_{dc1} at the rectifier side, which can be directly obtained in practice.

The estimated sliding surface of system (28) is defined as

$$\begin{bmatrix} \widehat{S}_{r1} \\ \widehat{S}_{r2} \end{bmatrix} = \begin{bmatrix} \hat{z}'_{11} - Q_1^* \\ \rho_1(\hat{z}_{11} - V_{dc1}^*) + \rho_2(\hat{z}_{12} - \dot{V}_{dc1}^*) \end{bmatrix} \tag{37}$$

where ρ_1 and ρ_2 are the positive sliding surface gains. The attractiveness of the estimated sliding surface (33) ensures reactive power Q_1 and DC voltage V_{dc1} can track to their reference.

The POSMC of system (28) is designed as

$$\begin{bmatrix} u_{d1} \\ u_{q1} \end{bmatrix} = B_{r0}^{-1} \begin{bmatrix} -\widehat{\Psi}_{r1}(\cdot) + \dot{Q}_1^* - \zeta'_r\widehat{S}_{r1} - \varphi'_r\mathrm{sat}\left(\widehat{S}_{r1}\right) \\ -\widehat{\Psi}_{r2}(\cdot) + \ddot{V}_{dc1}^* - \rho_1(\hat{z}_{12} - \dot{V}_{dc1}^* - \zeta_r\widehat{S}_{r2} - \varphi_r\mathrm{sat}\left(\widehat{S}_{r2}\right) \end{bmatrix} \tag{38}$$

where positive control gains ζ_r, ζ'_r, φ_r, ρ_1, and φ'_r are chosen to ensure the attractiveness of estimated sliding surface (33).

During the most severe disturbance, both the reactive power and DC voltage reduce from their initial value to around zero within a short period of time Δ. Thus, the boundary values of the system state and perturbation estimates can be obtained as $|\hat{z}'_{11}| \le |Q_1^*|$, $|\widehat{\Psi}_{r1}(\cdot)| \le |Q_1^*|/\Delta$, $|\hat{z}_{11}| \le |V_{dc1}^*|$, $|\hat{z}_{12}| \le |V_{dc1}^*|/\Delta$, and $|\widehat{\Psi}_{r2}(\cdot)| \le |V_{dc1}^*|/\Delta^2$, respectively.

3.3. Inverter controller design

Choose the system output $y_i = [y_{i1}, y_{i2}]^T = [Q_2, P_2]^T$, let Q_2^* and P_2^* be the given references of the reactive and active power, respectively. Define the tracking error $e_i = [e_{i1}, e_{i2}]^T = [Q_2 - Q_2^*, P_2 - P_2^*]^T$, differentiate e_i for inverter (2) until the control input appears explicitly, yields

$$\begin{bmatrix} \dot{e}_{i1} \\ \dot{e}_{i2} \end{bmatrix} = \begin{bmatrix} f_{i1} - \dot{Q}_2^* \\ f_{i2} - \dot{P}_2^* \end{bmatrix} + B_i \begin{bmatrix} u_{d2} \\ u_{q2} \end{bmatrix} \tag{39}$$

where

$$\begin{cases} f_{i1} &= \dfrac{3u_{sq2}}{2}\left(-\dfrac{R_2}{L_2} i_{d2} + \omega i_{q2}\right) \\ f_{i2} &= \dfrac{3u_{sq2}}{2}\left(-\dfrac{R_2}{L_2} i_{q2} - \omega i_{d2}\right) \end{cases} \tag{40}$$

and

$$B_i = \begin{bmatrix} \dfrac{3u_{sq2}}{2L_2} & 0 \\ 0 & \dfrac{3u_{sq2}}{2L_2} \end{bmatrix} \tag{41}$$

The determinant of matrix B_i is obtained as $|B_i| = 9u_{s2}^2/(4L_2^2)$, which is nonzero within the operation range of the inverter, thus system (35) is linearizable.

Assume all the nonlinearities are unknown, define the perturbations $\Psi_{i1}(\cdot)$ and $\Psi_{i2}(\cdot)$ as

$$\begin{bmatrix} \Psi_{i1}(\cdot) \\ \Psi_{i2}(\cdot) \end{bmatrix} = \begin{bmatrix} f_{i1} \\ f_{i2} \end{bmatrix} + (B_i - B_{i0}) \begin{bmatrix} u_{d2} \\ u_{q2} \end{bmatrix} \tag{42}$$

where the constant control gain B_{i0} is given by

$$B_{i0} = \begin{bmatrix} b_{i10} & 0 \\ 0 & b_{i20} \end{bmatrix} \tag{43}$$

Then system (35) can be rewritten as

$$\begin{bmatrix} \dot{e}_{i1} \\ \dot{e}_{i2} \end{bmatrix} = \begin{bmatrix} \Psi_{i1}(\cdot) \\ \Psi_{i2}(\cdot) \end{bmatrix} + B_{i0} \begin{bmatrix} u_{d2} \\ u_{q2} \end{bmatrix} - \begin{bmatrix} \dot{Q}_2^* \\ \dot{P}_2^* \end{bmatrix} \tag{44}$$

Similarly, define $z'_{21} = Q_2$ and $z_{21} = P_2$, two second-order SMPOs are used to estimate $\Psi_{i1}(\cdot)$ and $\Psi_{i2}(\cdot)$, respectively, as

$$\begin{cases} \dot{\hat{z}}'_{21} = \widehat{\Psi}_{i1}(\cdot) + \alpha'_{i1}\tilde{Q}_2 + k'_{i1}\mathrm{sat}(\tilde{Q}_2) + b_{i10}u_{d2} \\ \dot{\widehat{\Psi}}_{i1}(\cdot) = \alpha'_{i2}\tilde{Q}_2 + k'_{i2}\mathrm{sat}(\tilde{Q}_2) \end{cases} \tag{45}$$

where observer gains k'_{i1}, k'_{i2}, α'_{i1}, and α'_{i2} are all positive constants.

$$\begin{cases} \dot{\hat{z}}_{21} = \widehat{\Psi}_{i2}(\cdot) + \alpha_{i1}\tilde{P}_2 + k_{i1}\mathrm{sat}(\tilde{P}_2) + b_{i20}u_{q2} \\ \dot{\widehat{\Psi}}_{i2}(\cdot) = \alpha_{i2}\tilde{P}_2 + k_{i2}\mathrm{sat}(\tilde{P}_2) \end{cases} \tag{46}$$

where observer gains k_{i1}, k_{i2}, α_{i1}, and α_{i2} are all positive constants.

The above observers (38) and (39) only need the measurement of reactive power Q_2 and active power P_2 at the inverter side, which can be directly obtained in practice.

The estimated sliding surface of system (35) is defined as

$$\begin{bmatrix} \widehat{S}_{i1} \\ \widehat{S}_{i2} \end{bmatrix} = \begin{bmatrix} \widehat{z}'_{21} - Q_2^* \\ \widehat{z}_{21} - P_2^* \end{bmatrix} \tag{47}$$

Similarly, the attractiveness of the estimated sliding surface (40) ensures the reactive power Q_2 and active power P_2 can track to their reference.

The POSMC of system (35) is designed as

$$\begin{bmatrix} u_{d2} \\ u_{q2} \end{bmatrix} = B_{i0}^{-1} \begin{bmatrix} -\widehat{\Psi}_{i1}(\cdot) + \dot{Q}_2^* - \zeta'_i\widehat{S}_{i1} - \varphi'_i\mathrm{sat}\left(\widehat{S}_{i1}\right) \\ -\widehat{\Psi}_{i2}(\cdot) + \dot{P}_2^* - \zeta_i\widehat{S}_{i2} - \varphi_i\mathrm{sat}\left(\widehat{S}_{i2}\right) \end{bmatrix} \tag{48}$$

where positive control gains ζ_i, ζ_i, φ_i, and φ'_i are chosen to ensure the attractiveness of estimated sliding surface (40).

Similarly, the boundary values of the system state and perturbation estimates can be obtained as $|\widehat{z}'_{21}| \le |Q_2^*|$, $|\widehat{\Psi}_{i1}(\cdot)| \le |Q_2^*|/\Delta$, $|\widehat{z}_{21}| \le |P_2^*|$, and $|\widehat{\Psi}_{i2}(\cdot)| \le |P_2^*|/\Delta$, respectively.

Note that control outputs (34) and (41) are modulated by the sinusoidal pulse width modulation (SPWM) technique [6] in this chapter. The overall controller structure of the VSC-HVDC system is illustrated by **Figure 2**, in which only reactive power Q_1 and DC voltage V_{dc1} need to be measured for rectifier controller (34), while active power P_2 and reactive power Q_2 for inverter controller (41).

Remark 3 The conventional linear PI/PID control scheme employs an inner current loop to regulate the current [11], which could employ a synchronous reference frame (SRF)-based current controller [43] to avoid overcurrent. In contrast, the proposed POSMC (34) and (41) actually contains no current in its control law, while it cannot handle the overcurrent. Hence, the overcurrent protection devices [44] will be activated to prevent the overcurrent to grow, which can be seen in **Figure 2**.

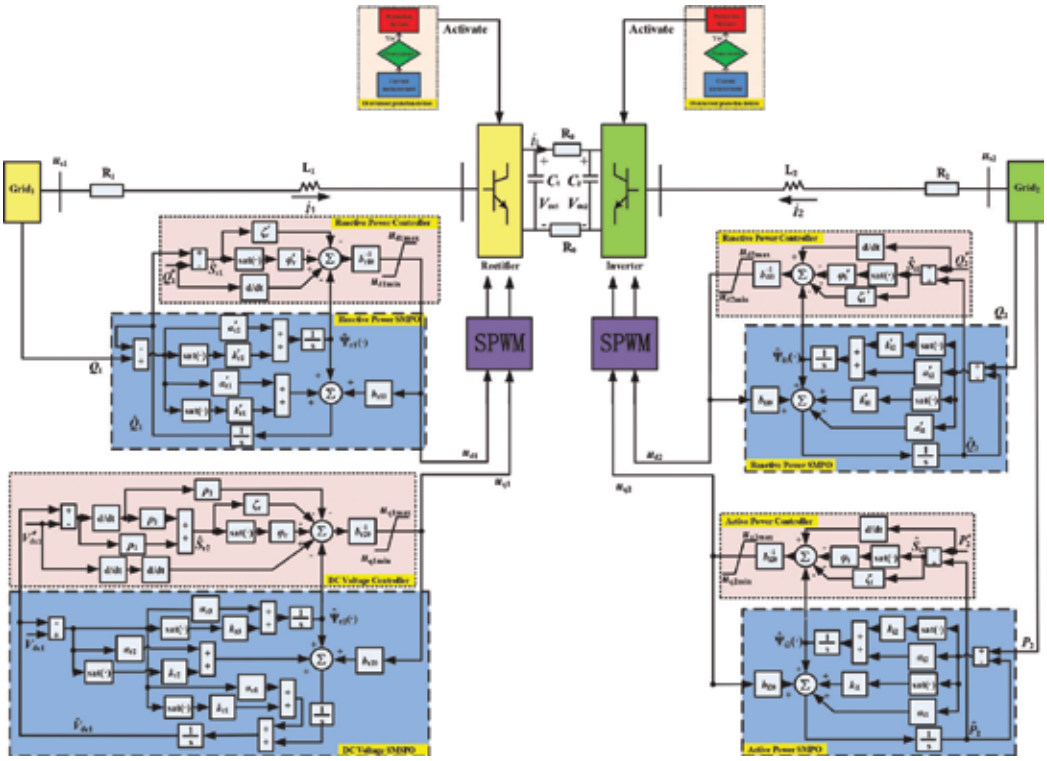

Figure 2. The overall controller structure of the VSC-HVDC system.

4. Simulation results

POSMC is applied on the VSC-HVDC system illustrated in **Figure 1**. The AC grid frequency is 50 Hz and VSC-HVDC system parameters are given in **Table 1**. POSMC parameters are provided in **Table 2**, in which the observer poles are allocated as $\lambda_{\alpha_r} = 100$ and $\lambda_{\alpha'_r} = \lambda_{\alpha_i} = \lambda_{\alpha'_i} = 20$,

AC system-based voltage	$V_{\text{AC base}}$	132 kV
DC cable base voltage	$V_{\text{DC base}}$	150 kV
System base power	S_{base}	100 MVA
AC system resistance (25 km)	R_1, R_2	0.05 Ω/km
AC system inductance (25 km)	L_1, L_2	0.026 mH/km
DC cable resistance (50 km)	R_0	0.21 Ω/km
DC bus capacitance	C_1, C_2	11.94 μF

Table 1. The VSC-HVDC system parameters.

Rectifier controller gains			
$b_{r10} = 100$	$b_{r20} = 7000$	$\rho_1 = 800$	$\rho_2 = 1$
$\zeta_r = 20$	$\zeta'_r = 10$	$\varphi_r = 20$	$\varphi'_r = 20$
Rectifier observer gains			
$\alpha_{r1} = 300$	$\alpha'_{r1} = 40$	$\alpha_{r2} = 3 \times 10^4$	$\alpha'_{r2} = 400$
$\alpha_{r3} = 10^6$	$\Delta = 0.01$	$\epsilon = 0.1$	$k_{r1} = 100$
$k'_{r1} = 75$	$k_{r2} = 10^5$	$k'_{r2} = 3.75 \times 10^4$	$k_{r3} = 2.5 \times 10^7$
Inverter controller gains			
$b_{i10} = 50$	$b_{i20} = 50$	$\zeta_i = 10$	$\zeta'_i = 10$
$\varphi_i = 10$	$\varphi'_i = 10$		
Inverter observer gains			
$\alpha_{i1} = 40$	$\alpha'_{i1} = 40$	$\alpha_{i2} = 400$	$\alpha'_{i2} = 400$
$k_{i1} = 75$	$k'_{i1} = 75$	$k_{i2} = 3.75 \times 10^4$	$k'_{i2} = 3.75 \times 10^4$

Table 2. POSMC parameters for the VSC-HVDC system.

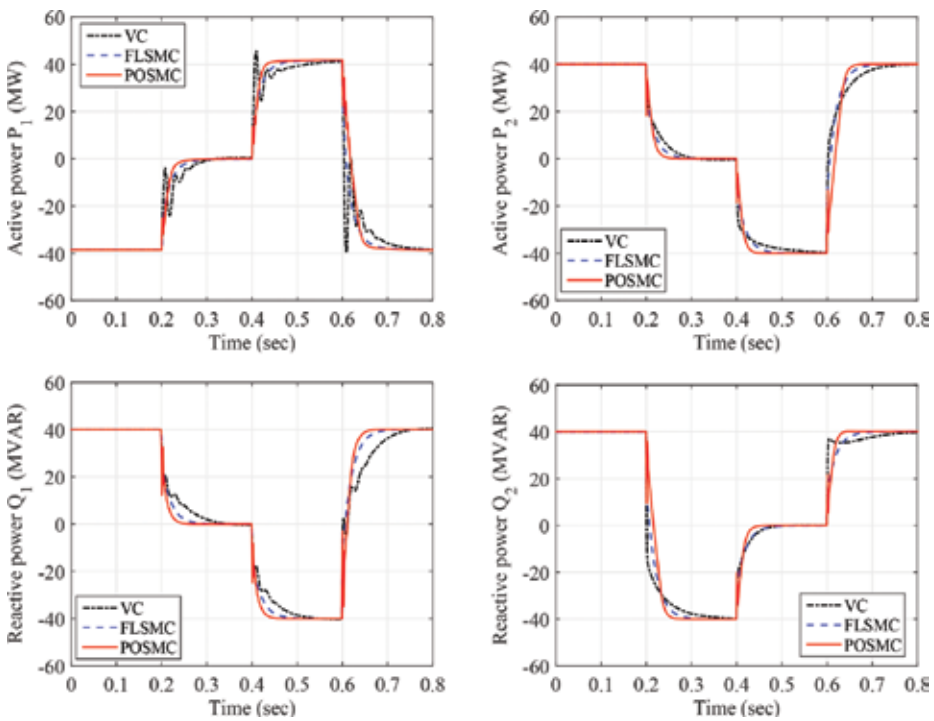

Figure 3. System responses obtained under the active and reactive power tracking.

while control inputs are bounded as $|u_{qi}| \leq 80$ kV and $|u_{di}| \leq 60$ kV, where $i = 1, 2$. The switching frequency is 1620 Hz for both rectifier and inverter, which is taken from [22]. The control performance of POSMC is compared to that of VC [11] and FLSMC [22] by the following four cases. In addition, two identical three-level neutral-point-clamped VSCs model for each rectifier and inverter from Matlab/Simulink SimPowerSystems are employed, which structure and parameters are taken directly from [11]. The simulation is executed on Matlab/Simulink 7.10 using a personal computer with an IntelR CoreTMi7 CPU at 2.2 GHz and 8 GB of RAM.

(1) Case 1: Active and reactive power tracking: The references of active and reactive power are set to be a series of step change occurs at $t = 0.2$ s, $t = 0.4$ s and restores to the original value at $t = 0.6$ s, while DC voltage is regulated at the rated value $V^*_{dc1} = 150$ kV. The system responses are illustrated by **Figure 3**. One can find that POSMC has the fastest tracking rate and maintains a consistent control performance under different operation conditions.

(2) Case 2: 5-cycle line-line-line-ground (LLLG) fault at AC bus 1: A five-cycle LLLG fault occurs at AC bus 1 when $t = 0.1$ s. Due to the fault, AC voltage at the corresponding bus is decreased to a critical level. **Figure 4** shows that POSMC can effectively restore the system with the smallest active power oscillations. Response of perturbation estimation is demonstrated in **Figure 5**, which shows that SMSPO and SMPO can estimate the perturbations with a fast tracking rate.

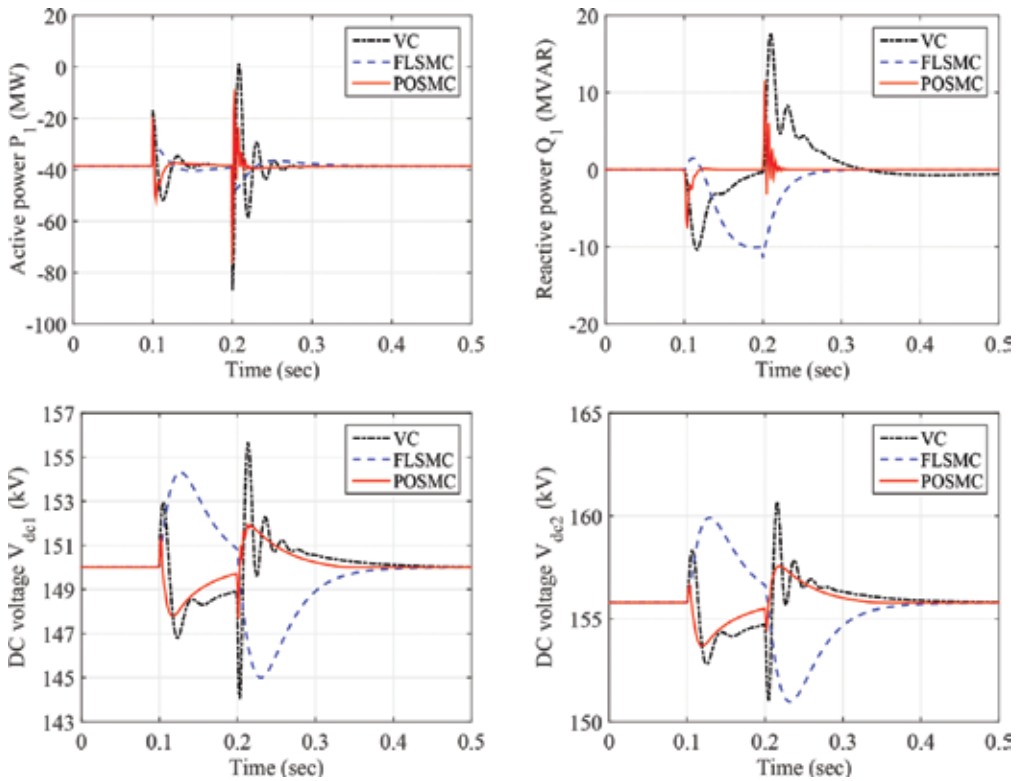

Figure 4. System responses obtained under the five-cycle LLLG fault at AC bus 1.

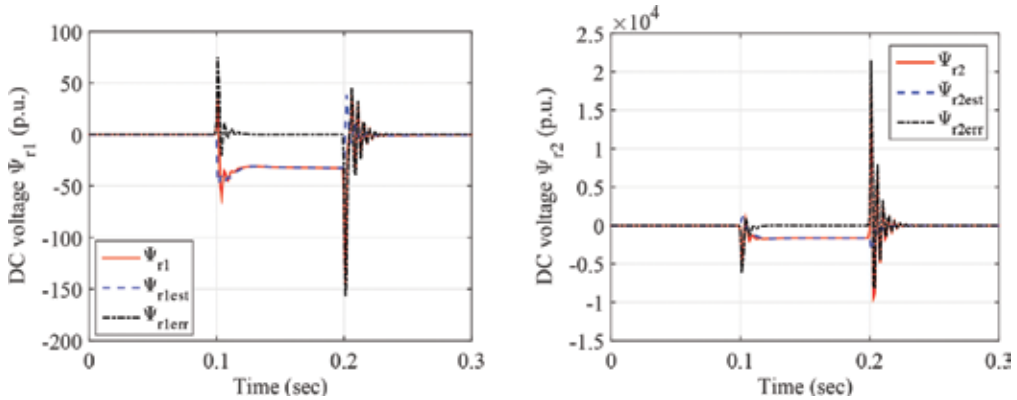

Figure 5. Estimation errors of the perturbations obtained under the five-cycle LLLG fault at AC bus 1.

(3) Case 3: Weak AC grid connection: The AC grids are assumed to be sufficiently strong such that AC bus voltages are ideal constants. It is worth considering a weak AC grid connected to the rectifier, e.g., offshore wind farms, which voltage u_{s1} is no longer a constant but a time-varying function. A voltage fluctuation that occurs from 0.15 to 1.05 s caused by the wind speed variation is applied, which corresponds to $u_{s1} = 1 + 0.15\sin(0.2\pi t)$. System responses are presented in **Figure 6**, it illustrates that both DC voltage and reactive power are oscillatory,

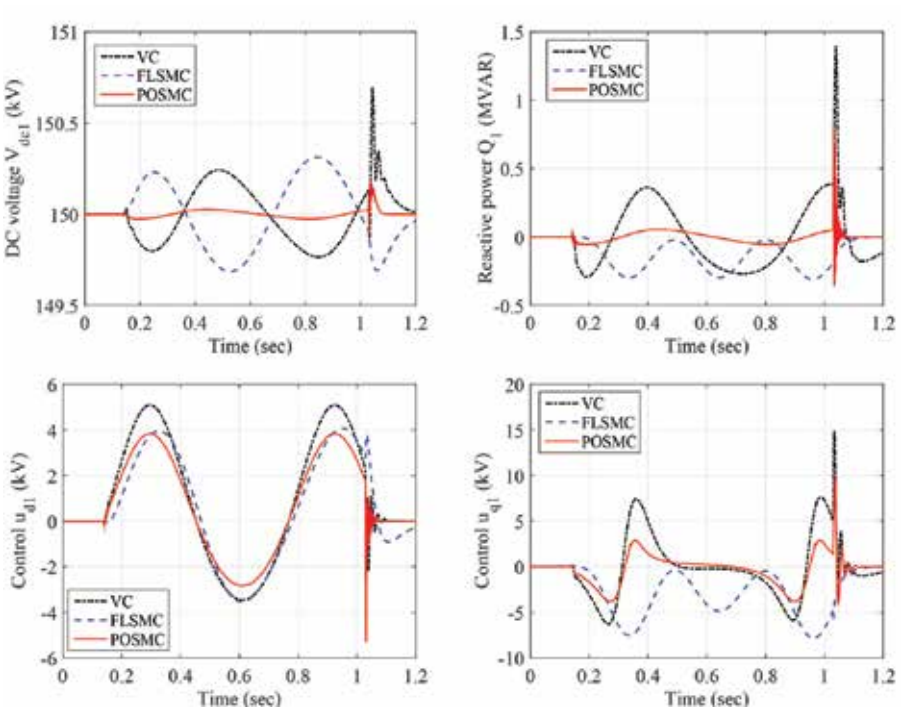

Figure 6. System responses obtained with the weak AC grid connection.

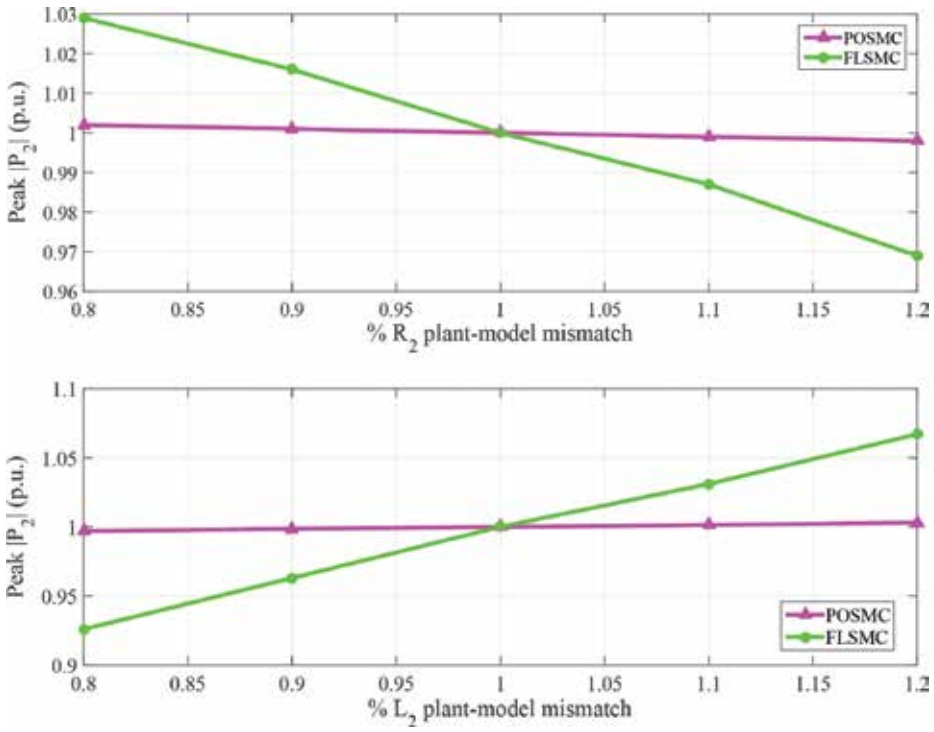

Figure 7. The peak active power $|P_2|$ (in p.u.) to a −120 A in the DC cable current i_L obtained at nominal grid voltage for plant-model mismatches in the range of 20% (one parameter changes and others keep constant).

while POSMC can effectively suppress such oscillation with the smallest fluctuation of DC voltage and reactive power.

(4) Case 4: System parameter uncertainties: When there is a fault in the transmission or distribution grid, the resistance and inductance values of the grid may change significantly. Several tests are performed for plant-model mismatches of R_2 and L_2 with $\pm 20\%$ uncertainties. All tests are undertaken under the nominal grid voltage and a corresponding −120 A in the DC cable current i_L at 0.1 s. The peak active power $|P_2|$ is recorded, which uses per unit (p.u.) value for a clear illustration of system robustness. It can be found from **Figure 7** that the peak active power $|P_2|$ controlled by POSMC is almost not affected, while FLSMC has a relatively large range of variation, i.e., around 3% to R_2 and 8% to L_2, respectively. Responses to mismatch of R_2 and L_2 changing at the same time are demonstrated in **Figure 8**. The magnitude of changes is around 10% under FLSMC and almost does not change under POSMC. This is because POSMC estimates all uncertainties and does not need an accurate system model, thus it has better robustness than that of FLSMC which requires accurate system parameters.

The integral of absolute error (IAE) indices of each approach calculated in different cases are tabulated in **Table 3**. Here, $\text{IAE}_{Q_1} = \int_0^T |Q_1 - Q_1^*| \mathrm{d}t$, $\text{IAE}_{V_{\text{dc1}}} = \int_0^T |V_{\text{dc1}} - V_{\text{dc1}}^*| \mathrm{d}t$, $\text{IAE}_{Q_2} = \int_0^T |Q_2 - Q_2^*| \mathrm{d}t$, and $\text{IAE}_{P_2} = \int_0^T |P_2 - P_2^*| \mathrm{d}t$. The simulation time T = 3 s. Note that POSMC has

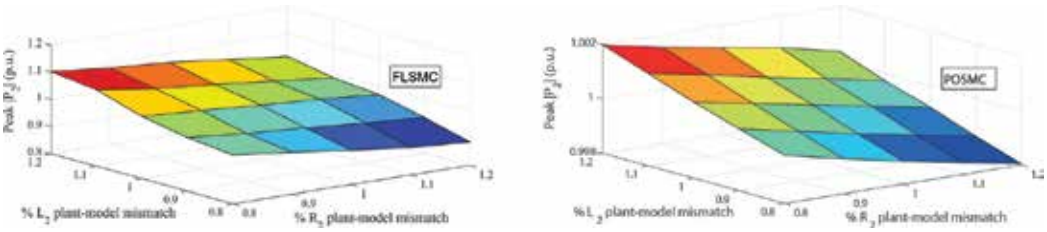

Figure 8. The peak active power $|P_2|$ (in p.u.) to a −120 A in the DC cable current i_L obtained at nominal grid voltage for plant-model mismatches in the range of 20% (different parameters may change at the same time).

Method	Power tracking			
Case				
	IAE_{Q_1}	$IAE_{V_{dc1}}$	IAE_{Q_2}	IAE_{P_2}
VC	3.83E-02	4.44E-03	2.13E-02	2.71E-02
FLSMC	2.19E-02	1.73E-03	2.23E-02	2.18E-02
POSMC	2.33E-02	2.00E-03	2.42E-02	2.33E-02
Method	Five-cycle LLLG fault		Weak AC grid connection	
Case				
	IAE_{Q_1}	$IAE_{V_{dc1}}$	IAE_{Q_1}	$IAE_{V_{dc1}}$
VC	2.62E-02	2.15E-03	4.53E-03	4.13E-03
FLSMC	1.13E-02	4.13E-03	4.08E-03	3.33E-03
POSMC	5.64E-03	1.38E-03	3.88E-04	6.78E-04

Table 3. IAE indices (in p.u.) of different control schemes calculated in different cases.

Figure 9. Overall control costs IAE_u (in p.u.) obtained in different cases.

a little bit higher IAE than that of FLSMC in the power tracking due to the estimation error, while it can provide much better robustness in the case of 5-cycle LLLG fault and weak AC grid connection. In particular, its IAE_{Q_1} and $\mathrm{IAE}_{V_{\mathrm{dc1}}}$ are only 8.57 and 9.51% of those of VC, 16.42 and 20.36% of those of FLSMC with the weak AC grid connection. The overall control costs are illustrated in **Figure 9**, with $\mathrm{IAE}_u = \int_0^T \left(|u_{\mathrm{d1}}| + |u_{\mathrm{q1}}| + |u_{\mathrm{d2}}| + |u_{\mathrm{q2}}|\right)\mathrm{d}t$. It is obvious that POSMC has the lowest control costs in all cases, which is resulted from the merits that the upper bound of perturbation is replaced by the smaller bound of its estimation error, thus an over-conservative control input can be avoided.

5. Hardware-in-the-loop test results

HIL test is an important and powerful technique used in the development and test of complex real-time embedded systems, which provides an effective platform by adding the complexity of the plant under control to the test platform. The complexity of the plant under control is included in test and development by adding a mathematical representation of all related dynamic systems.

A dSPACE simulator-based HIL test is used to validate the implementation feasibility of POSMC, which configuration and experiment platform are given by **Figures 10** and **11**, respectively. The rectifier controller (34) and inverter controller (41) are implemented on one dSPACE platform (DS1104 board) with a sampling frequency $f_{\mathrm{c}} = 1$ kHz, and the VSC-HVDC system is simulated on another dSPACE platform (DS1006 board) with the limit sampling frequency

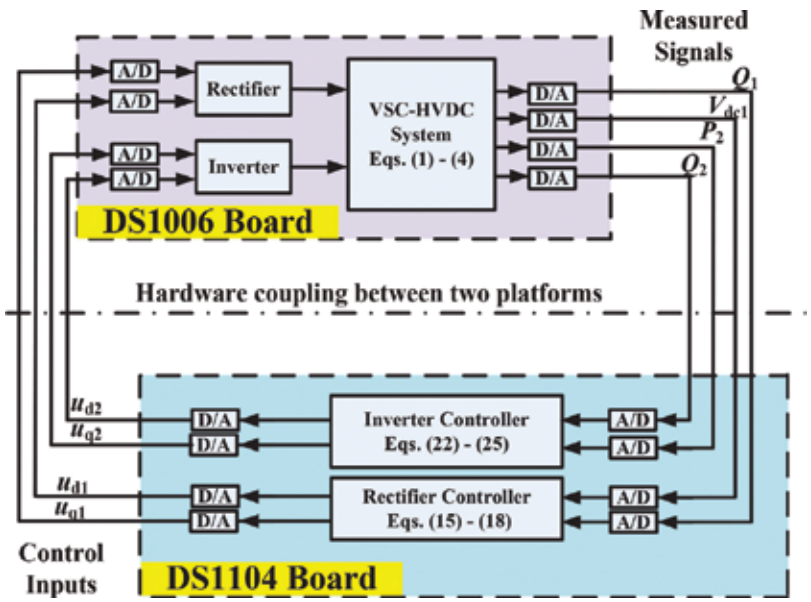

Figure 10. The configuration of the HIL test.

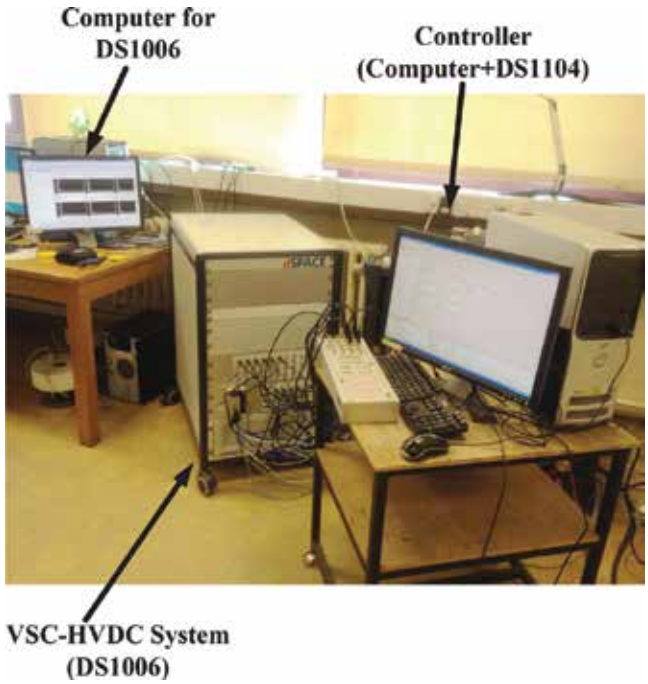

Figure 11. The experiment platform of the HIL test.

$f_s = 50$ kHz to make HIL simulator as close to the real plant as possible. The measurements of the reactive power Q_1, DC voltage V_{dc1}, active power P_2, and reactive power Q_2 are obtained from the real-time simulation of the VSC-HVDC system on the DS1006 board, which are sent to two controllers implemented on the DS1104 board for the control inputs calculation.

It follows from [37] that an unexpected high-frequency oscillation in control inputs may emerge as the large observer poles would result in high gains, which lead to highly sensitive observer dynamics to the measurement disturbances in the HIL test. Note that this phenomenon does not exist in the simulation. One effective way to alleviate such malignant effect is to reduce the observer poles. Through trial-and-error, an observer pole in the range of $\lambda_{\alpha_r} \in [15, 25]$ and $\lambda_{\alpha'_r} = \lambda_{\alpha_i} = \lambda_{\alpha'_i} \in [3, 10]$ can avoid such oscillation but with almost similar transient responses, thus the reduced poles $\lambda_{\alpha_r} = 20$ and $\lambda_{\alpha'_r} = \lambda_{\alpha_i} = \lambda_{\alpha'_i} = 5$, with $b_{r10} = 50$, $b_{r20} = 5000$, $b_{i10} = 20$, and $b_{i20} = 20$, are chosen in the HIL test. Furthermore, a time delay $\tau = 3$ ms has been assumed in the corresponding simulation to consider the effect of the computational delay of the real-time controller.

(1) Case 1: Active and reactive power tracking: The reference of active and reactive power changes at $t = 0.4$ s, $t = 0.9$ s and restores to the original value at $t = 1.4$ s, while DC voltage is regulated at the rated value $V^*_{dc1} = 150$ kV. The system responses obtained under the HIL test and simulation are compared by **Figure 12**, which shows that the HIL test has almost the same results as that of the simulation.

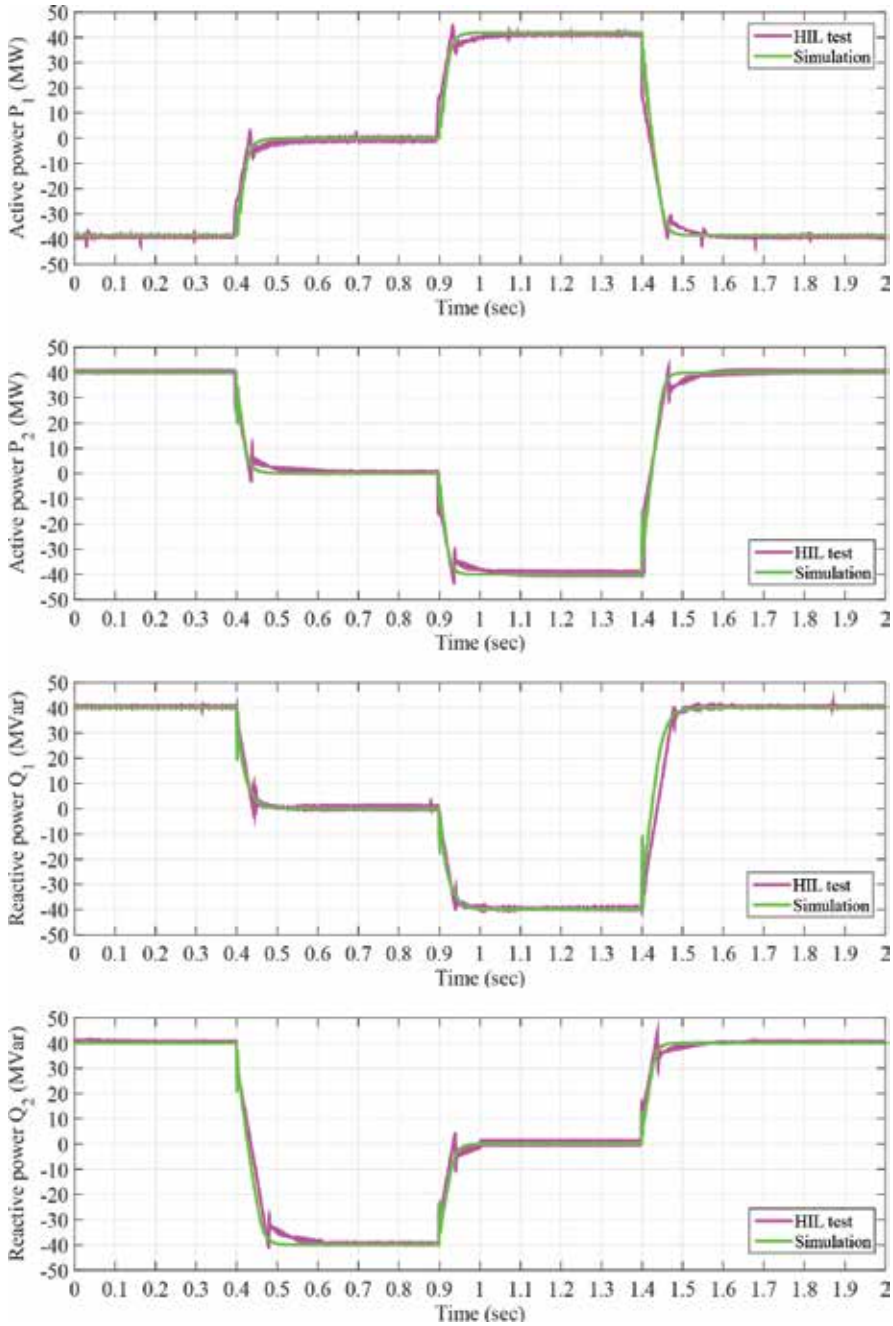

Figure 12. HIL test results of system responses obtained under the active and reactive power tracking.

(2) Case 2: 5-cycle line-line-line-ground (LLLG) fault at AC bus 1. A 5-cycle LLLG fault occurs at AC bus 1 when $t = 0.1$ s. **Figure 13** demonstrates that the system can be rapidly restored and the system responses obtained by the HIL test is similar to that of simulation.

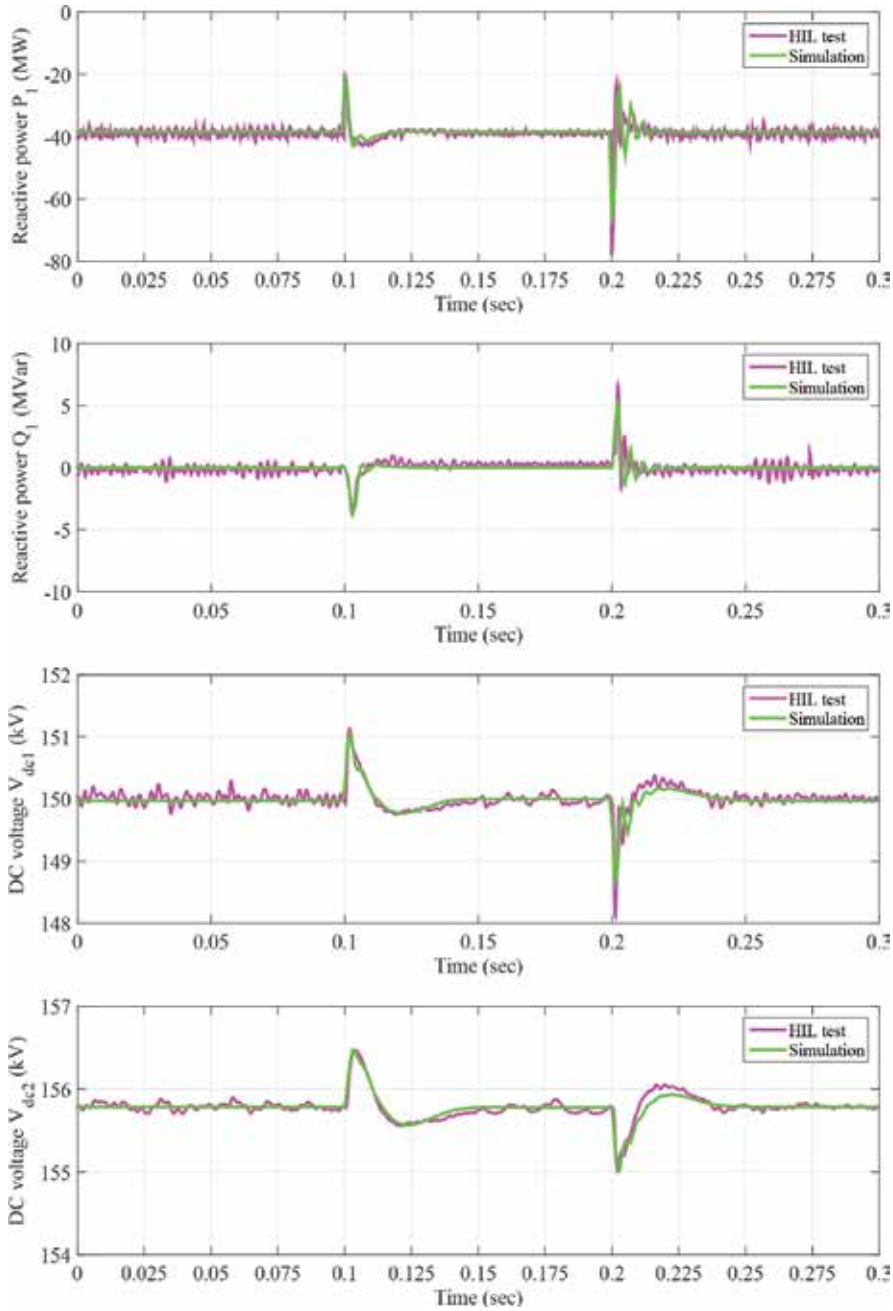

Figure 13. HIL test results of system responses obtained under the five-cycle LLLG fault at AC bus 1.

(3) Case 3: Weak AC grid connection: The same voltage variation $u_{s1} = 1 + 0.15\sin(0.2\pi t)$ is applied between 0.87 and 2.45 s. It can be readily seen from **Figure 14** that the results of the HIL test and simulation match very well.

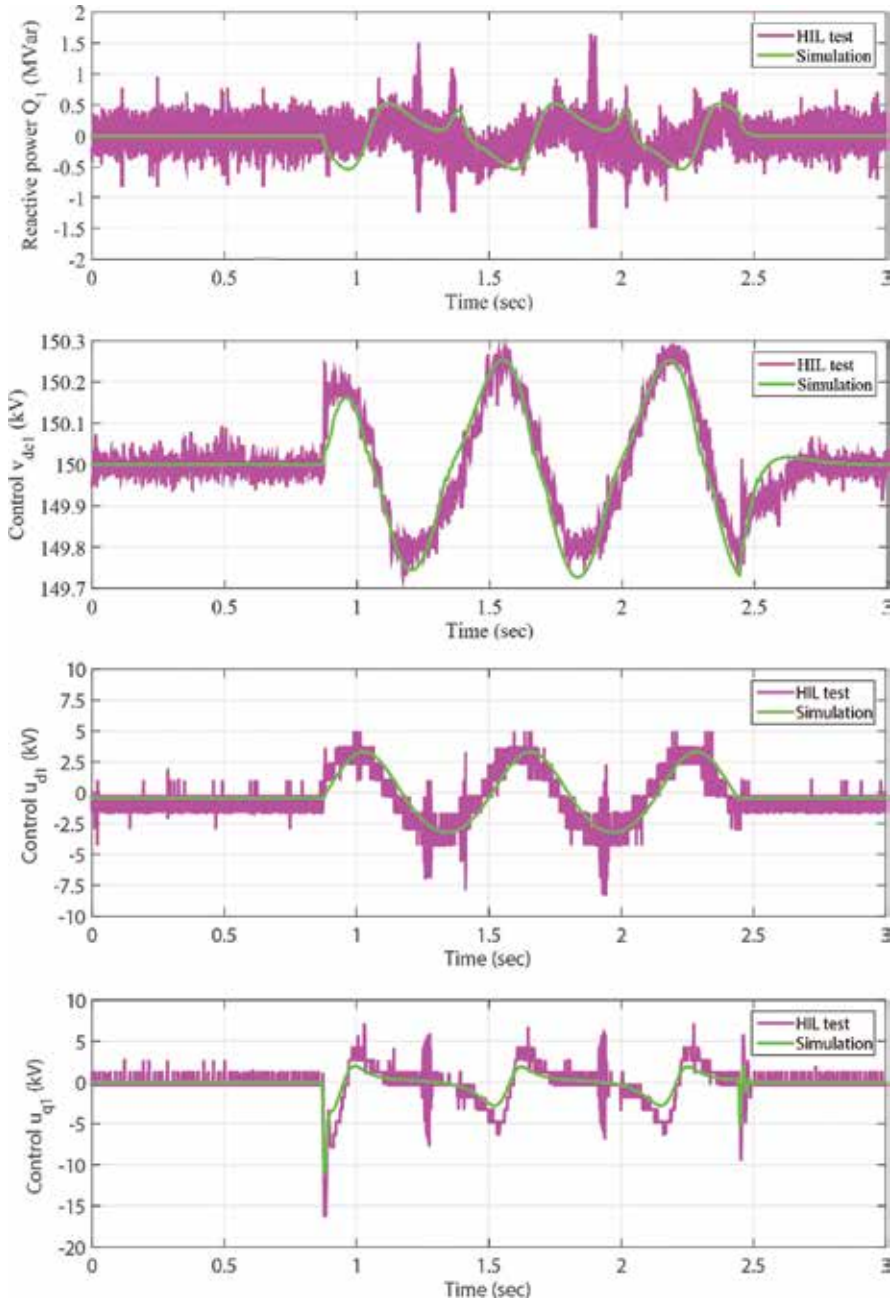

Figure 14. HIL test results of system responses obtained with the weak AC grid connection.

The difference of the obtained results between the HIL test and simulation is possibly due to the following two reasons:

- There exist measurement disturbances in the HIL test, which are, however, not taken into account in the simulation, a filter could be used to remove the measurement disturbances, thus the control performance can be improved.

- The sampling frequency of VSC-HVDC model and POSMC is the same in simulation ($f_s = f_c = 1$ kHz) as they are implemented in Matlab of the same computer. In contrast, the sampling frequency of VSC-HVDC model ($f_s = 50$ kHz) is significantly increased in the HIL test to make VSC-HVDC model as close to the real plant as possible. Note the sampling frequency of POSMC remains the same ($f_c = 1$ kHz) due to the sampling limit of the practical controller.

6. Conclusion

A POSMC scheme has been developed for the VSC-HVDC system to rapidly compensate the combinatorial effect of nonlinearities, parameter uncertainties, unmodeled dynamics, and time-varying external disturbances. As the upper bound of perturbation is replaced by the smaller bound of its estimation error, an over-conservative control input is avoided such that the tracking accuracy can be improved.

Four case studies have been undertaken to evaluate the control performance of the proposed approach, which verify that POSMC can maintain a consistent control performance with less power overshoot during the power reversal, restore the system rapidly after the AC fault, suppress the oscillation effectively when connected to a weak AC grid, and provide significant robustness in the presence of system parameter uncertainties. At last, a dSPACE-based HIL test has been carried out which validates the implementation feasibility of POSMC.

Acknowledgements

This work is supported by National Natural Science Foundation of China under Grant Nos. 51477055, 51667010, and 51777078.

Author details

Bo Yang[1]*, Tao Yu[2], Hongchun Shu[1] and Pulin Cao[1]

*Address all correspondence to: yangbo_ac@outlook.com

1 Faculty of Electric Power Engineering, Kunming University of Science and Technology, Kunming, China

2 School of Electrical Engineering, South China University of Technology, Guangzhou, Guangdong, China

References

[1] Yang B, Yu T, Shu HC, Dong J, Jiang L. Robust sliding-mode control of wind energy conversion systems for optimal power extraction via nonlinear perturbation observers. Applied Energy. 2018;**210**:711-723

[2] Yang B, Hu YL, Huang HY, Shu HC, Yu T, Jiang L. Perturbation estimation based robust state feedback control for grid connected DFIG wind energy conversion system. International Journal of Hydrogen Energy. 2017;**42**(33):20994-21005

[3] Yang B, Sang YY, Shi K, Yao W, Jiang L, Yu T. Design and real-time implementation of perturbation observer based sliding-mode control for VSC-HVDC systems. Control Engineering Practice. 2016;**56**:13-26

[4] Yang B, Yu T, Zhang XS, Huang LN, Shu HC, Jiang L. Interactive teaching-learning optimizer for parameter tuning of VSC-HVDC systems with offshore wind farm integration. IET Generation Transmission and Distribution. 2018;**12**:678-687

[5] Yang B, Jiang L, Yao W, Wu QH. Perturbation observer based adaptive passive control for damping improvement of multi-terminal voltage source converter-based high voltage direct current systems. Transactions of the Institute of Measurement and Control. 2017; **39**(9):1409-1420

[6] Flourentzou N, Agelidis VG, Demetriades GD. VSC-based HVDC power transmission systems: An overview. IEEE Transactions on Power Electronics. 2009;**24**(3):592-602

[7] Hertema DV, Ghandhari M. Multi-terminal VSC HVDC for the European supergrid: Obstacles. Renewable and Sustainable Energy Reviews. 2010;**14**:3156-3163

[8] Zhang L, Harnefors L, Nee HP. Interconnection of two very weak AC systems by VSC-HVDC links using power-synchronization control. IEEE Transactions on Power Systems. 2011;**26**(1):344-355

[9] Zhang L, Harnefors L, Nee HP. Modeling and control of VSC-HVDC links connected to island systems. IEEE Transactions on Power Systems. 2011;**26**(2):783-793

[10] Haileselassie TM, Molinas M, Undeland T. Multi-terminal VSCHVDC system for integration of offshore wind farms and green electrification of platforms in the North Sea. Presented at the Nordic Workshop on Power and Industrial Electronics, Otakaari, Finland. 2008

[11] Li S, Haskew TA, Xu L. Control of HVDC light system using conventional and direct current vector control approaches. IEEE Transactions on Power Electronics. 2010;**25**(12): 3106-3118

[12] Ruan SY, Li GJ, Peng L, Sun YZ, Lie TT. A nonlinear control for enhancing HVDC light transmission system stability. International Journal of Electrical Power & Energy Systems. 2007;**27**:565-570

[13] Durrant M, Werner H, Abbott K. Synthesis of multi-objective controllers for a VSC HVDC terminal using LMIs. In: IEEE Conference of Decision and Control; 2004. pp. 4473-4478

[14] Ruan SY, Li GJ, Jiao XH, Sun YZ, Lie T. Adaptive control design for VSC-HVDC systems based on backstepping method. Electric Power Systems Research. 2007;**77**:559-565

[15] Lordelo A, Fazzolari H. On interval goal programming switching surface robust design for integral sliding mode control. Control Engineering Practice. 2014;**32**:136-146

[16] Huo W. Predictive variable structure control of nonholonomic chained systems. International Journal of Computer Mathematics-Computer Mathematics in Dynamics and Control. 2008;**85**(6):949-960

[17] Zong Q, Zhao ZS, Zhang J. Higher order sliding mode control with self-tuning law based on integral sliding mode. IET Control Theory and Applications. 2010;**4**(7):1282-1289

[18] Gokasan M, Bogosyan S, Goering DJ. Sliding mode based powertrain control for efficiency improvement in series hybrid-electric vehicles. IEEE Transactions on Power Electronics. 2006;**21**(3):779-790

[19] Kessal A, Rahmani L. Ga-optimized parameters of sliding-mode controller based on both output voltage and input current with an application in the PFC of AC/DC converters. IEEE Transactions on Power Electronics. 2014;**29**(6):3159-3165

[20] Lascu C, Boldea I, Blaabjerg F. Direct torque control of sensorless induction motor dirves: A sliding-mode approach. IEEE Transactions on Industrial Electronics. 2004;**40**(2):582-590

[21] Beltran B, Ahmedali T, Benbouzid MEH. Sliding mode power control of variable-speed wind energy conversion systems. Electric Machines & Drives Conference, 2007. IEMDC '07. IEEE International, Vol. 23, IEEE; pp. 551-558. 2008

[22] Moharana A, Dash PK. Input-output linearization and robust sliding-mode controller for the VSC-HVDC transmission link. IEEE Transactions on Power Delivery. 2010;**25**(3):1952-1961

[23] Edwards C, Spurgeon S. Sliding Mode Control: Theory and Applications. London, UK: CRC Press; 1998

[24] Johnson C. Accommodation of external disturbances in linear regulator and servomechanism problems. IEEE Transactions on Automatic Control. 1971;**16**(6):635-644

[25] Chen WH, Ballance DJ, Gawthrop PJ, O'Reilly J. A nonlinear disturbance observer for robotic manipulators. IEEE Transactions on Industrial Electronics. 2000;**27**(4):932-938

[26] She JH, Fang M, Ohyama Y, Hashimoto H, Wu M. Improving disturbance-rejection performance based on an equivalent-input disturbance approach. IEEE Transactions on Industrial Electronics. 2008;**55**(1):380-389

[27] Sun L, Li DH, Lee KY. Enhanced decentralized PI control for fluidized bed combustor via advanced disturbance observer. Control Engineering Practice. 2015;**42**:128-139

[28] Han JQ. From PID to active disturbance rejection control. IEEE Transactions on Industrial Electronics. 2009;**56**:900-906

[29] Sun L, Dong JY, Li DH, Lee KY. A practical multivariable control approach based on inverted decoupling and decentralized active disturbance rejection control. Industrial & Engineering Chemistry Research. 2016;**55**(7):2008-2019

[30] Guo BZ, Zhao ZL. On the convergence of an extended state observer for nonlinear systems with uncertainty. Systems and Control Letters. 2011;**60**(6):420-430

[31] Kwon SJ, Chung WK. Perturbation Compensator Based Robust Tracking Control and State Estimation of Mechanical Systems. New York: Springer; 2004

[32] Xia Y, Zhu Z, Fu M. Back-stepping sliding mode control for missile systems based on an extended state observer. IET Control Theory and Applications. 2011;**5**(1):93-102

[33] Yue M, Liu BY, An C, Sun XJ. Extended state observer-based adaptive hierarchical sliding mode control for longitudinal movement of a spherical robot. Nonlinear Dynamics. 2014; **78**:1233-1244

[34] Wang JX, Li SH, Yang J, Wu B, Li Q. Extended state observer-based sliding mode control for PWM-based DC-DC buck power converter systems with mismatched disturbances. IET Control Theory and Applications. 2015;**9**(4):579-586

[35] Jiang L, Wu QH, Wen JY. Nonlinear adaptive control via sliding-mode state and perturbation observer. IEE Proceedings - Control Theory and Applications. 2002;**149**(4): 269-277

[36] Liu Y, Wu QH, Zhou XX, Jiang L. Perturbation observer based multiloop control for the DFIGWT in multimachine power system. IEEE Transactions on Power Systems. 2014; **29**(6):2905-2915

[37] Yang B, Jiang L, Yao W, Wu QH. Perturbation estimation based coordinated adaptive passive control for multimachine power systems. Control Engineering Practice. 2015;**44**: 172-192

[38] Jovcic D. Phase locked loop system for FACTS. IEEE Transactions on Power Systems. 2003;**18**(3):1116-1124

[39] Jiang L.Nonlinear adaptive control and applications in power systems [PhD thesis]. University of Liverpool; 2001

[40] Hernandez J, Barbot JP. Sliding observer-based feedback control for flexible joints manipulator. Automatica. 1996;**32**(9):1243-1254

[41] Khalil HK. Nonlinear Systems. 3rd ed. New Jersey: Prentice Hall; 2002

[42] Slotine JJE, Li W. Applied Nonlinear Control. London: Prentice-Hall; 1991

[43] Geddada N, Mishra MK, Kumar MV. SRF based current controller using PI and HC regulators for DSTATCOM with SPWM switching. International Journal of Electrical Power & Energy Systems. 2015;**67**:87-100

[44] Baran ME, Mahajan NR. Overcurrent protection on voltage-source-converter-based multiterminal DC distribution systems. IEEE Transactions on Power Delivery. 2007;**22**:406-412

Chapter 3

A Formal Perturbation Theory of Carleman Operators

Sidi Mohamed Bahri

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.79022

Abstract

In this chapter, we introduce a multiplication operation that allows us to give to the Carleman integral operator of second class the form of a multiplication operator. Also we establish the formal theory of perturbation of such operators.

Keywords: Carleman kernel, defect indices, integral operator, formal series

1. Introduction

In this chapter, we shall assume that the reader is familiar with the fundamental results and the standard notation of the integral operators theory [1–3, 5, 6, 8–12]. Let X be an arbitrary set, μ be a $\sigma-$ finite measure on X (μ is defined on a $\sigma-$ algebra of subsets of X; we do not indicate this $\sigma-$ algebra), and $L_2(X,\mu)$ the Hilbert space of square integrable functions with respect to μ. Instead of writing "$\mu-$ measurable," "$\mu-$ almost everywhere," and "$(d\mu(x))$," we write "measurable," "a e," and "dx."

A linear operator A: $D(A) \longrightarrow L_2(X,\mu)$, where the domain $D(A)$ is a dense linear manifold in $L_2(X,\mu)$, is said to be integral if there exist a measurable function K on $X \times X$, a kernel, such that, for every $f \in D(A)$,

$$Af(x) = \int_X K(x,y)f(y)dy \quad \text{a.e..} \tag{1}$$

A kernel K on $X \times X$ is said to be Carleman, if $K(x,y) \in L_2(X,\mu)$ for almost every fixed x, that is to say

$$\int_X |K(x,y)|^2 dy \ < \infty \quad \text{a.e..} \tag{2}$$

An integral operator A (1) with a kernel K is called Carleman operator, if K is a Carleman kernel (2). Every Carleman kernel K defines a Carleman function k from X to $L_2(X,\mu)$ by $k(x) = \overline{K(x,.)}$ for all x in X for which $K(x,.) \in L_2(X,\mu)$..

Now we consider the Carleman integral operator (1) of second class [3, 8] generated by the following symmetric kernel:

$$K(x,y) = \sum_{p=0}^{\infty} a_p \psi_p(x) \overline{\psi_p(y)}, \tag{3}$$

where the overbar in (3) denotes the complex conjugation and $\left\{\psi_p(x)\right\}_{p=0}^{\infty}$ is an orthonormal sequence in $L^2(X,\mu)$ such that

$$\sum_{p=0}^{\infty} \left|\psi_p(x)\right|^2 < \infty \ \text{ a.e.,} \tag{4}$$

and $\{a_p\}_{p=0}^{\infty}$ is a real number sequence verifying

$$\sum_{p=0}^{\infty} a_p^2 \left|\psi_p(x)\right|^2 < \infty \text{a.e..} \tag{5}$$

We call $\left\{\psi_p(x)\right\}_{p=0}^{\infty}$ a Carleman sequence.

Moreover, we assume that there exist a numeric sequence $\left\{\gamma_p\right\}_{p=0}^{\infty}$ such that

$$\sum_{p=0}^{\infty} \gamma_p \psi_p(x) = 0 \ \text{ a.e.,} \tag{6}$$

and

$$\sum_{p=0}^{\infty} \left|\frac{\gamma_p}{a_p - \lambda}\right|^2 < \infty. \tag{7}$$

With the conditions (6) and (7), the symmetric operator $A = (A^*)^*$ admits the defect indices $(1,1)$ (see [3]), and its adjoint operator is given by

$$A^* f(x) = \sum_{p=0}^{\infty} a_p \left(f, \psi_p\right) \psi_p(x), \tag{8}$$

$$D(A^*) = \left\{ f \in L^2(X,\mu) : \sum_{p=0}^{\infty} a_p \left(f, \psi_p\right) \psi_p(x) \in L^2(X,\mu) \right\}. \tag{9}$$

Moreover, we have

$$\begin{cases} \varphi_\lambda(x) = \sum_{p=0}^{\infty} \frac{\gamma_p}{a_p - \lambda} \psi_p(x) \in \mathfrak{N}_{\overline{\lambda}}, \quad \lambda \in \mathbb{C}, \quad \lambda \neq a_k, \quad k = 1, 2, \ldots \\ \varphi_{a_k}(x) = \psi_k(x), \end{cases} \tag{10}$$

$\mathfrak{N}\overline{\lambda}$ being the defect space associated with λ (see [3, 4])..

2. Position operator

Let $\psi = \{\psi_n\}_{n=0}^{\infty}$ be a fixed Carleman sequence in $L^2(X, \mu)$. It is clear from the foregoing that ψ is not a complete sequence in $L^2(X, \mu)$. We denote by $\mathfrak{L}_\psi$ the closure of the linear span of the sequence $\left\{\psi_p(x)\right\}_{p=0}^{\infty}$:

$$\mathfrak{L}_\psi = \overline{span\{\psi_n, n \in \mathbb{N}\}}. \tag{11}$$

We start this section by defining some formal spaces.

2.1. Formal elements

Definition 1. *(see* [7]*) We call formal element any expression of the form*

$$f = \sum_{n \in \mathbb{N}} a_n \psi_n, \tag{12}$$

where the coefficients $a_n (n \in \mathbb{N})$ are scalars.

The sequence $(a_n)_n$ is said to generate the formal element f.

Definition 2. *We say that f is the zero formal element, and we note $f = 0$ if $a_n = 0$ for all $n \in \mathbb{N}$.*

We say that two formal elements $f = \sum_{n \in \mathbb{N}} a_n \psi_n$ and $g = \sum_{n \in \mathbb{N}} b_n \psi_n$ are equal if $a_n = b_n$ for all $n \in \mathbb{N}$.

If φ is a scalar function defined for each a_n, we set

$$\varphi\left(\sum_n a_n \psi_n\right) = \sum_n \varphi(a_n) \psi_n, \tag{13}$$

or in another form,

$$\varphi(a_1, a_2, \ldots, a_n, \ldots) = (\varphi(a_1), \varphi(a_2), \ldots, \varphi(a_n), \ldots). \tag{14}$$

For example, let

$$\varphi(x) = \frac{1}{x}, \quad (x \neq 0). \tag{15}$$

If $a_n \neq 0$ for all $n \in \mathbb{N}$, then the formal element

$$\varphi\left(\sum_n a_n \psi_n\right) = \sum_n \frac{1}{a_n} \psi_n \tag{16}$$

is called inverse of the formal element $f = \sum_n a_n \psi_n$.

Furthermore, we define the conjugate of a formal element f by

$$\overline{f} = \sum_n \overline{a_n} \psi_n. \tag{17}$$

Denote by $\mathcal{F}_\psi$ the set of all formal elements (12).

On $\mathcal{F}_\psi$, we define the following algebraic operations:

the sum

$$\begin{array}{rcccl} + & : & \mathcal{F}_\psi \times \mathcal{F}_\psi & \to & \mathcal{F}_\psi \\ & & \left(\sum_n a_n \psi_n\right) + \left(\sum_n b_n \psi_n\right) & = & \sum_n (a_n + b_n) \psi_n \end{array} \tag{18}$$

and the product

$$\begin{array}{rcccl} \cdot & : & \mathbb{C} \times \mathcal{F}_\psi & \to & \mathcal{F}_\psi \\ & & \lambda \cdot \left(\sum_n a_n \psi_n\right) & = & \sum_n (\lambda . a_n) \psi_n. \end{array} \tag{19}$$

Hence, we obtain a complex vector space structure for $\mathcal{F}_\psi$.

2.2. Bounded formal elements

Definition 3. *A formal element $f = \sum_{n \in \mathbb{N}} a_n \psi_n$ is bounded if its sequence $(a_n)_n$ is bounded.*

We denote by $\mathcal{B}_\psi$ the set of all bounded formal elements.

It is clear that $\mathcal{B}_\psi$ is a subspace of $\mathcal{F}_\psi$.

We claim that:

1. $\mathcal{L}_\psi$ is a subspace of $\mathcal{B}_\psi$.

2. Furthermore we have the strict inclusions:

$$\mathcal{L}_\psi \subset \mathcal{B}_\psi \subset \mathcal{F}_\psi. \tag{20}$$

We define a linear form $\langle .,. \rangle$ on $\mathcal{F}_\psi$ by setting:

$$\left\langle \sum_n a_n \psi_n, \sum_n b_n \psi_n \right\rangle = \sum_n a_n \overline{b_n} \tag{21}$$

with the series converging on the right side of (21).

Proposition 4. *The form (21) verifies the properties of scalar product.*

Proof. Indeed, let

$$f = \sum_n a_n \psi_n, g = \sum_n b_n \psi_n, f_1 = \sum_n a_n^1 \psi_n \text{and } f_2 = \sum_n a_n^2 \psi_n$$

in $\mathcal{F}_\psi$.

We have then:

1. $\langle f, g \rangle = \sum_n a_n \overline{b_n} = \overline{\sum_n a_n \overline{b_n}} = \overline{\langle f, g \rangle}$.

2.
$$\langle \lambda f, g \rangle = \left\langle \lambda \left(\sum_n a_n \psi_n \right), \sum_n a_n \psi_n \right\rangle = \left\langle \sum_n (\lambda a_n) \psi_n, \sum_n b_n \psi_n \right\rangle$$
$$= \sum_n (\lambda a_n) \overline{b_n} = \lambda \left\langle \sum_n a_n \psi_n, \sum_n b_n \psi_n \right\rangle = \lambda \langle f, g \rangle.$$

3.
$$\langle f_1 + f_2, g \rangle = \left\langle \sum_n \left(a_n^1 + a_n^2 \right) \psi_n, \sum_n b_n \psi_n \right\rangle$$
$$= \sum_n \left(a_n^1 + a_n^2 \right) \overline{b_n} = \sum_n a_n^1 \overline{b_n} + \sum_n a_n^2 \overline{b_n} = \langle f_1, g \rangle + \langle f_2, g \rangle.$$

4. $\langle f, f \rangle = \sum_n |a_n|^2 \geq 0 \text{and} \langle f, f \rangle > 0 \text{ if } f \neq 0.$

■

Remark 5. *On $\mathcal{L}_\psi$, the scalar product $\langle .,. \rangle$ coincides with the scalar product $(.,.)$ of $L^2(X, \mu)$.*

2.3. The multiplication operation

Here, we introduce the crucial tool of our work.

Definition 6. *We call multiplication with respect to the Carleman sequence $\{\psi_n\}_n$, the operation denoted "$\circ$" and defined by:*

$$f \circ g = \sum_n \langle f, \psi_n \rangle \langle g, \psi_n \rangle = \sum_n a_n b_n \psi_n, \quad \forall (f, g) \in \mathcal{F}_\psi^2. \tag{22}$$

Definition 7. *We call position operator in $\mathcal{L}_\psi$ any unitary self-adjoint (see [1]) operator satisfying*

$$U(f \circ g) = (Uf \circ Ug), \quad \text{for all } f, g \in \mathcal{L}_\psi. \tag{23}$$

The term "position operator" comes from the fact (as it will be shown in the following theorem) that for the elements of the sequence $\psi = \{\psi_n\}_n$, the operator U acts as operator of change of position of these elements.

2.4. Main results

Theorem 8. *A linear operator defined on $\mathcal{L}_\psi$ is a position operator if and only if there exist an involution j (i.e., $j^2 = Id$) of the set $\mathbb{N}$ such that for all $n \in \mathbb{N}$*

$$U\psi_n = \psi_{j(n)}. \tag{24}$$

Proof.

1. It is easy to see that if (24) holds, then U is a position operator.

2. Let U be a position operator. According to 1, we can write

$$U\psi_n = \sum_k \alpha_{n,k} \ \psi_k \ \ with \sum_k |\alpha_{n,k}|^2 = 1 \tag{25}$$

since $U\psi_n \in \mathcal{L}_\psi$..

On the other hand, we have

$$\sum_k \alpha_{n,k} \ \psi_k = \sum_k \alpha_{n,k}^2 \ \psi_k \tag{26}$$

as

$$U\psi_n = U(\psi_n \circ \psi_n) = U\psi_n \circ U\psi_n.$$

The equalities (26) lead to the resolution of the system:

$$\begin{cases} \sum_n \alpha_{n,k}^2 = 1, \\ \alpha_{n,k}^2 = \alpha_{n,k}, \quad k \in \mathbb{N}. \end{cases} \tag{27}$$

We get then

$$(\forall n \in \mathbb{N})\ (\exists! k_n \in \mathbb{N}) : \begin{cases} \alpha_{n,k_n} = 1, \\ \alpha_{n,k} = 0 \qquad \forall k \neq k_n. \end{cases}$$

Let us now consider the following application:

$$\begin{array}{rccl} j & : & \mathbb{N} & \to & \mathbb{N}, \\ & & n & \mapsto & j(n) = k_n. \end{array}$$

It's clear that j is injective.

Now let $m \in \mathbb{N}$. Since $U^2 = I$, then

$$U(U\psi_m) = U\psi_{j(m)} = \psi_{j(j(m))} = \psi_m.$$

Hence,

$$j(j(m)) = m.$$

Finally j is well defined as involution.

■

Notation In the sequel, $j(n)$ will be noted by n^v. We write

$$U\psi_n = \psi_{j(n)} = \psi_n^v \tag{28}$$

and

$$Uf = U\left(\sum_n a_n \ \psi_n\right) = \sum_n a_n \ \psi_n^v = \overset{v}{f} \tag{29}$$

Remark 9. *The position operator U can be extended over $\mathcal{F}_\psi$ as follows:*

If $f = \sum_n a_n \ \psi_n \in \mathcal{F}_\psi$, then

$$Uf = \overset{v}{f} = \sum_n a_n \ \psi_n^v. \tag{30}$$

3. Carleman operator in $\mathcal{F}_\psi$

3.1. Case of defect indices $(1,1)$

Let $\alpha = \sum_p \alpha_p \psi_p \in \mathcal{F}_\psi$; we introduce the operator A°_α} defined in $\mathcal{L}_\psi$ by

$$\mathring{A}_\alpha f = \alpha \circ f = \sum_n \langle \alpha, \psi_n \rangle \langle f, \psi_n \rangle \psi_n. \tag{31}$$

It is clear that $\mathring{A}_\alpha$ is a Carleman operator induced by the kernel

$$k(x,y) = \sum \alpha_n \psi_n(x) \overline{\psi_n(y)}, \tag{32}$$

with domain

$$D\left(\mathring{A}_\alpha\right) = \left\{ f \in \mathcal{L}_\psi : \sum_n \left|\alpha_n (f, \psi_n)\right|^2 < \infty \right\}. \tag{33}$$

Moreover, if $\alpha = \overline{\alpha}$, $\mathring{A}_\alpha$ is self-adjoint.

Now *let* $\Theta = \sum_p \gamma_p \psi_p \in \mathcal{F}_\psi$ and $\Theta \notin \mathcal{L}_\psi \left(i.e., \quad \sum_p \left|\gamma_p\right|^2 = \infty \right)$. We introduce the following set

$$\mathcal{H}_\Theta = \left\{ f + \mu\Theta : f \in \mathcal{L}_\psi, \mu \in \mathbb{C} \right\} \tag{34}$$

which verifies the following properties.

Proposition 10. *1.* $\mathcal{H}_\Theta$ *is a subset of* $\mathcal{F}_\psi$.

2. $\mathcal{H}_\theta = \mathcal{L}_\psi \oplus \mathbb{C}\theta$, i.e., direct sum of $\mathcal{L}_\psi$ with $\mathbb{C}\theta = \left\{ \mu\theta : \mu \in \mathbb{C} \right\}$.

Proof. The first property is easy to establish. We show the uniqueness for the second.

Let $g_1 = f_1 + \mu_1\theta$ and $g_2 = f_2 + \mu_2\theta$, two formal elements in $\mathcal{H}_\theta$. Then

$$g_1 = g_2 \Leftrightarrow f_1 - f_2 = (\mu_2 - \mu_1)\theta.$$

This last equality is verified only if $\mu_2 = \mu_1$. Therefore, $f_1 = f_2$. ■

Denote by Q the projector of $\mathcal{H}_\Theta$ on $\mathcal{L}_\psi$, that is to say: if $g \in \mathcal{H}_\Theta$,

$$g = f + \mu\Theta \text{ with } f \in \mathcal{L}_\psi \text{ and } \mu \in \mathbb{C}$$

then

$$Qg = f.$$

We define the operator B_α by:

$$B_\alpha f = Q(\alpha \circ f), f \in \mathcal{L}_\psi. \tag{35}$$

It is clear that

$$D(B_\alpha) = \left\{ f \in \mathcal{L}_\psi : (\alpha \circ f) \in \mathcal{H}_\Theta \right\}. \tag{36}$$

Theorem 11 B_α *is a densely defined and closed operator.*

Proof.

1. Since

$$span\left\{ \psi_n, n \in \mathbb{N} \right\} \subset D(B_\alpha)$$

and that $\left\{ \psi_n \right\}_n$ is complete in $\mathcal{L}_\psi$, then

$$\overline{D(B_\alpha)} = \mathcal{L}_\psi.$$

2. Let $(f_n)_n$ be a sequence of elements in $D(B_\alpha)$. Checking:

$$\begin{cases} f_n & \to & f \\ B_\alpha f_n & \to & g \end{cases} (\text{convergence in the } L^2 \text{ sense}).$$

We have then

$$B_\alpha f_n = Q(\alpha \circ f_n),$$

with

$$\alpha \circ f_n = g_n + \mu\Theta, g_n \in \mathcal{L}_\psi.$$

Then

$$g_n = \alpha \circ f_n - \mu_n\Theta \in \mathcal{L}_\psi,$$

This implies that

$$\langle g_n, \psi_m \rangle = \alpha_m \langle f_n, \psi_m \rangle - \mu_n \gamma_m \psi_m \quad \forall m \in \mathbb{N}.$$

Or, when n tends to ∞, we have

$$g_n \to g \text{ and } f_n \to f.$$

Therefore, there exist $\mu \in \mathbb{C}$ such that

$$\lim_{n\to\infty} \mu_n = \mu.$$

And as Q is a closed operator, then we can write

$$(\alpha \circ f) \in \mathcal{H}_\Theta \text{ and } g = Q(\alpha \circ f).$$

Finally $f \in D(B_\alpha)$ and $g = B_\alpha f$.

■

It follows from this theorem that the adjoint operator B_α^* exists and $B_\alpha^{**} = B_\alpha$.

Let us denote by A_α the operator adjoint of B_α,

$$A_\alpha = B_\alpha^*. \tag{37}$$

In the case $\alpha = \overline{\alpha}$, the operator A_α is symmetric and we have the following results:

Theorem 12. *A_α admits defect indices $(1,1)$ if and only if*

$$\varphi_\lambda = (\alpha - \lambda)^{-1} \circ \Theta \in \mathcal{L}_\psi. \tag{38}$$

In this case $\varphi_\lambda \in \mathcal{N}\overline{\lambda}$ *(defect space associated with* λ*,* [3]).

Proof. We know (see [3]) that A_α has the defect indices $(1,1)$ if and only if its defect subspaces $\mathcal{N}\overline{\lambda}$ and $\mathcal{N}_\lambda$ are unidimensional.

We have

$$\mathcal{N}_{\overline{\lambda}} = \ker\left(A_\alpha^* - \lambda I\right) = \ker(B_\alpha - \lambda I).$$

So it suffices to solve the system:

$$\begin{cases} B_\alpha \varphi_\lambda = \lambda \varphi_\lambda \\ \varphi_\lambda \in \mathcal{L}_\psi \end{cases}$$

that is,

$$\begin{cases} Q(\alpha \circ \varphi_\lambda) = \lambda \varphi_\lambda \\ \varphi_\lambda \in \mathcal{L}_\psi \end{cases} \Leftrightarrow \begin{cases} (\alpha \circ \varphi_\lambda) = \lambda \varphi_\lambda + \mu\Theta, \mu \in \mathbb{C} \\ \varphi_\lambda \in \mathcal{L}_\psi \end{cases}$$

$$\Leftrightarrow \begin{cases} (\alpha - \lambda) \circ \varphi_\lambda = \Theta \\ \varphi_\lambda \in \mathcal{L}_\psi \end{cases}$$

$$\Leftrightarrow \begin{cases} \varphi_\lambda = (\alpha - \lambda)^{-1} \circ \Theta \\ \varphi_\lambda \in \mathcal{L}_\psi \end{cases}.$$

■

3.2. Case of defect indices $(\mathbf{m}, \mathbf{m})$

In this section, we give the generalization for the case of defect indices (m, m), $m > 1$.

Let $\Theta_1, \Theta_2, \ldots, \Theta_m$, m be formal elements not belonging to $\mathcal{L}_\psi$, and let

$$\mathcal{H}_\Theta = \left\{ f + \sum_{k=1}^{m} \mu_k \Theta_k, \quad f \in \mathcal{L}_\psi, \mu_k \in \mathbb{C}, \quad k = 1, \ldots, m \right\}. \tag{39}$$

We consider the operator B_α defined by

$$\begin{aligned} B_\alpha f &= Q(\alpha \circ f) \quad f \in D(B_\alpha), \\ D(B_\alpha) &= \left\{ f \in \mathcal{L}_\psi : \alpha \circ f \in \mathcal{H}_\Theta \right\} \end{aligned} \tag{40}$$

We assume that $\alpha = \overline{\alpha}$ and we set

$$A_\alpha = B_\alpha^*. \tag{41}$$

By analogy to the case of defect indices $(1,1)$, we also have the following:

Theorem 13. *The operator B_α is densely defined and closed.*

Theorem 14. *The operator A_α admits defect indices (m, m) if and only if*

$$\varphi_\lambda^{(k)} = (\alpha - \lambda) \circ \Theta_k \in \mathcal{L}_\psi, k = 1, \ldots, m. \tag{42}$$

In this case, the functions $\varphi_\lambda^{(k)} (k = 1, \ldots, m)$ are linearly independent and generate the defect space $\mathcal{N}\overline{\lambda}$.

4. Conclusion

We have seen the interest of multiplication operators in reducing Carleman integral operators and how they simplify the spectral study of these operators with some perturbation. In the same way, we can easily generalize this perturbation theory to the case of the non-densely defined Carleman operators:

$$\begin{aligned} H(x,y) = K(x,y) + \sum_{j=1}^{m} b_j \psi_j(x) \varphi_j(y), \\ \left(\varphi_j \in L^2(X,\mu), \psi_j \notin L^2(X,\mu), j = \overline{1,m}\right), \end{aligned} \tag{43}$$

with $K(x, y)$ a Carleman kernel.

It should be noted that this study allows the estimation of random variables.

Author details

Sidi Mohamed Bahri

Address all correspondence to: sidimohamed.bahri@univ-mosta.dz

Laboratory of Pure and Applied Mathematics, Abdelhamid Ibn Badis University, Mostaganem, Algeria

References

[1] Akhiezer NI, Glazman IM. Theory of Linear Operators in Hilbert Space. New York: Dover; 1993

[2] Aleksandrov EL, On the resolvents of symmetric operators which are not densely defined, Izvestiya Vysshikh Uchebnykh Zavedenii. Matematika, 1970 N 7, 3–12

[3] Bahri SM. On the extension of a certain class of Carleman operators, EMS. ZAA. 2007;**26**(1): 57-64

[4] Bahri SM. Spectral properties of a certain class of Carleman operators, EMS. Archivum Mathematicum (Brno). 2007;**3**:43

[5] Bahri SM. On convex hull of orthogonal scalar spectral functions of a Carleman operator. Boletim da Sociedade Paranaense de Matemática. 2008;**26**(1-2):9-18

[6] Bahri SM, On the spectra of quasi self adjoint extensions of a Carleman operator, Mathematica Bohemica, 2012 Vol. 137, N 3

[7] Belbahri, Kamel M. Series de laurent formelles. Office des Publications Universitaires, Alger; 1981

[8] Carleman T. Sur les équations intégrales singulières à noyau réel et symétrique. Uppsala Lundequistska bokhandeln; 1923

[9] Korotkov VB. Integral operators with Carleman kernels (in Russian). Nauka, Novosibirsk; 1983

[10] Targonski GI. On Carleman integral operators. Proceedings of the American Mathematical Society. 1967;**18**(3):450-456

[11] Weidman J. Carleman Operators. Manuscripta Mathematica. 1970;**2**:1-38

[12] Weidman J. Linear Operators in Hilbert Spaces. New-York Heidelberg Berlin: Spring Verlag; 1980

Chapter 4

On Optimal and Simultaneous Stochastic Perturbations with Application to Estimation of High-Dimensional Matrix and Data Assimilation in High-Dimensional Systems

Hong Son Hoang and Remy Baraille

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.77273

Abstract

This chapter is devoted to different types of optimal perturbations (OP), deterministic, stochastic, OP in an invariant subspace, and simultaneous stochastic perturbations (SSP). The definitions of OPs are given. It will be shown how the OPs are important for the study on the predictability of behavior of system dynamics, generating ensemble forecasts as well as in the design of a stable filter. A variety of algorithm-based SSP methodology for estimation and decomposition of very high-dimensional (Hd) matrices are presented. Numerical experiments will be presented to illustrate the efficiency and benefice of the perturbation technique.

Keywords: predictability, optimal perturbation, invariant subspace, simultaneous stochastic perturbation, dynamical system, filter stability, estimation of high-dimensional matrix

1. Introduction

Study in high-dimensional systems (HdS) today constitutes one of the most important research subjects thanks to the exponential increase of the power and speed of computers: after Moore's law, the number of transistors in a dense integrated circuit doubles approximately every 2 years (see Myhrvold [1]). However, this exponential increase is still far from being sufficient for responding to great demand on computational and memory resources in implementing the

optimal data assimilation algorithms (like Kalman filter (KF) [2], for example) for operational forecasting systems (OFS).

This chapter is devoted to the role of perturbations as an efficient tool for predictability of dynamical system, ensemble forecasting and for overcoming the difficulties in the design of data assimilation algorithms, in particular, of the optimal adaptive filtering for extremely HdS. In [3, 4], Lorenz has studied the problem of predictability of the atmosphere. It is found that the atmosphere is a chaotic system and a predictability limit to numerical forecast is of about 2 weeks. The barrier of predictability has to be overcome in order to increase the time period of a forecast further. The fact that estimates of the current state are inaccurate and that numerical models have inadequacies leads to forecast errors that grow with increasing forecast lead time. Ensemble forecasting aims at quantifying this flow-dependent forecast uncertainty. Today, a medium-range forecast has become a standard product. In the 1990s, the ensemble forecasting (EnF) technique was introduced in operational centers such as the European Centre for Medium-range Weather Forecast (ECMWF) (see Palmer et al. [5]), the NCEP (US National Center for Environmental Prediction) (Toth and Kalnay [6]). It is found that a single forecast can depart rapidly from the real atmosphere. The idea of the ensemble forecasting is to add the perturbations around the control forecast to produce a collection of forecasts that try to better simulate the possible uncertainties in a numerical forecast. The ensemble mean can then act as a nonlinear filter such that its skill is higher than that of individual members in a statistical sense (Toth and Kalnay [6]).

The chapter is organized as follows. Section 2 outlines first the optimal perturbation (OP) theory, on how the OP plays the important role for seeking the most growing direction of prediction error (PE). The predictability theory of the dynamical system as well as a stability of the filtering algorithm all are developed on the basis of OP. The definition of the optimal deterministic perturbation (ODP) and some theoretical results on the ODP are introduced. It is found that the ODP is associated with the right singular vector (SV) of the system dynamics. In Section 3, the two other classes of ODPs are presented: the leading eigenvector (EV) and real Schur vector (SchV) of the system dynamics. Mention that the first EV is the ODS in the eigen invariant subspace (EI-InS) of the system dynamics. As to the leading SchV, it is ODS in the Schur invariant subspace (Sch-InS) which is closely related to the EI-InS in the sense that the subspace of the leading SchVs, generated by the sampling procedure (Sampling-P, Section 3), converges to the EI-InS. In Section 4, we present the other type of OP called as optimal stochastic perturbation (OSP). Mention that the OSP is a natural extension of the ODP which gives insight into understanding of what represents the most growing PE and how one can produce it by stochastically perturbing the initial state. One important class of perturbations (known as simultaneous stochastic perturbation—SSP) is presented in Section 5. It will be shown that the SSP is very efficient for solving optimization problems in high-dimensional (Hd) setting. The different algorithms for estimating, decomposing … Hd matrices are also presented here. Numerical examples are presented in Section 6 for illustrating the theoretical results and efficiency of the OPs in solving data assimilation problems. The experiment on data assimilation in the Hd ocean model MICOM by the filters constructed on the basis of the Schur ODSs and SSPs is presented in Section 7. The concluding remarks are presented in Section 8.

http://dx.doi.org/10.5772/intechopen.77273

2. Optimal perturbations: predictability and filter stability

2.1. Stability of filter

The behavior of atmosphere or ocean is recognized as highly sensitive to initial conditions. It means that a small change in an initial condition can alter strongly the trajectory of the system. It is therefore important to be able to know about the directions of rapid growth of the system state. The research on OP is namely aimed at finding the methods to better capture these rapidly growing directions of the system dynamics, to optimize the predictability of the physical process under consideration.

To explain this phenomenon more clearly, consider a standard linear filtering problem

$$x(k+1) = \Phi x(k) + w(k+1), z(k+1) = Hx(k+1) + v(k+1). \tag{1}$$

where $\Phi \in R^{n\times n}$ is the state transition matrix, $H \in R^{p\times n}$ is an observation matrix. Under standard conditions related to the model and observation noises w_k, v_k, the minimum mean squared (MMS) estimate $\widehat{x}_k$ can be obtained by the well-known KF [2].

$$\widehat{x}(k+1) = \widehat{x}(k+1/k) + K\zeta(k+1), \widehat{x}(k+1/k) = \Phi\widehat{x}(k) \tag{2}$$

where $\zeta(k+1) = z(k+1) - H\widehat{x}(k+1/k)$ is the innovation vector, $\widehat{x}(k+1)$ is the filtered (or analysis) estimate, $\widehat{x}(k+1/k)$ is the one-step ahead prediction for $x(k+1)$. The KF gain K is given by

$$K = MH^T\left[HMH^T + R\right]^{-1} \tag{3}$$

From Eq. (2), it can be shown that the transition matrix for the filtered estimate equation is expressed by $L = [I - KH]\Phi$.

For HdS, the KF gain (3) is impossible to compute. In a study by Hoang et al. [7], it is suggested to find the gain with the structure

$$K = P_r K_e \tag{4}$$

with $P_r \in R^{n\times n_e}$—an operator projecting a vector from the reduced space R^{n_e} to the full system space R^n, $K_e \in R^{n_e\times p}$ is the gain for the reduced filter. One of very important questions arising here is how one can choose a subspace of projection and structure of K_e to make L to be stable? It is found in the work done by Hoang et al. [8] that detectability of the input-output system (1) is sufficient for the existence of a stabilizing gain K and this gain can be constructed with P_r consisting from all unstable EVs (or unstable SVs, SchVs. See Section 3) of the system dynamics.

2.2. Singular value decomposition and optimal perturbations

Consider the singular value decomposition (SVD) of Φ [9],

$$\begin{aligned}&\Phi = UDV^T, D = \text{diag}\,[\sigma_1, \sigma_2, \ldots, \sigma_n], \sigma_1 \geq \sigma_2 \ldots \geq \sigma_n,\\ &U = [U_1, U_2], D = \text{block diag}\ \ [D_1, D_2], V = [V_1, V_2]\end{aligned} \tag{5}$$

where $U_1, V_1 \in R^{\times n_1}$, $D_{n_1} \in R^{n_1 \times n_1}$, n_1 is the number of all unstable and neutral SVs of Φ. In the future, for simplicity, unless otherwise stated, we say on the set of all unstable SVs as that including all unstable and neutral SVs.

Suppose the system is *s-detectable* (detectability of all the columns of U_1 or V_1). From the research by Hoang et al. [10], there exists a stabilizing gain K_s (sufficient but not necessary condition) with $P_r = U_1$ (see Eq. (4)) such that the transition matrix $L = (I - K_s H)\Phi$ is stable. It signifies that the filter is stable and the estimation error is bounded. The columns of U_1, i.e., the left unstable SVs of Φ, serve as a basis for seeking appropriate correction in the filtering algorithm.

On the other hand, in practice, for extreme HdS, one cannot compute all elements of U_1 but only some of its subset $U_1' \in U_1$. Using U_1' instead of U_1 cannot guarantee a filter stability. The ensemble forecasting has been proposed as an approach to prevent a possible large error in the forecast and requires a knowledge on the rapidly growing directions of the PE. In this context, the OPs appear to be important which allow to search the rapid growing directions of the PE by model integration of OPs. They (i.e., OPs) are infact the unstable right SVs (RSV).

2.3. Optimal perturbation

Let δx be a given perturbation, representing an error (uncertainty, deterministic, or stochastic) around the true system state x^*, i.e., $\widehat{x}_f = x^* + \delta x$, (at some instant k). The prediction of the system state $\widehat{x}_p$ can be obtained by forwarding the numerical model (1) on the basis of $\widehat{x}_f$—filtered estimate, i.e., $\widehat{x}_p = \Phi\widehat{x}_f$. We have then

$$\widehat{x}_p = \Phi\widehat{x}_f = \Phi[x^* + \delta x] = \Phi x^* + \Phi\delta x, \tag{6}$$

One sees the perturbation δx in the initial system state "grows" into $\Phi\delta x$ which represents uncertainty in the forecast.

In general, the perturbation δx may be any element in the n-dimensional space, R^n, i.e., $\delta x \in R^n$. For $e_f^{(l)} := \delta x_f^{(l)}$—a sample of the filtered error (FE) e_f, integrating the model by e_f results in $e_p^{(l)} := \Phi\delta x_f^{(l)}$—a sample for the PE e_p. By generating the ensemble of perturbations $E_f(L) := \left\{e_f^{(l)}, l = 1, \ldots, L\right\}$ according to the distribution of e_f, one can produce the ensemble of PE samples $E_p(L) := \left\{e_p^{(l)}, l = 1, \ldots, L\right\}$ and use them to estimate the distribution of the PE e_p. This serves as a basis for the particle filtering [11].

The ensemble-based filtering (EnBF) algorithm is simplified for the standard linear filtering problems with e_f being of zero mean and the error covariance matrix (ECM) P. This technique is aimed to approximate the ECM without solving the matrix Ricatti equation (Ensemble KF—EnKF, [12]. Mention that at the present, one can generate only about O(100) samples at each

assimilation instant. This ensemble size is too small compared to the system dimension. That is why it is important to have a good strategy for selecting the "optimal" samples (perturbations) to better approximate the ECM in the filtering algorithm.

2.3.1. Optimal deterministic perturbation

Introduce

$$\left\{S(\delta x) : \delta x, \|\delta x\|_2 =< \delta x, \delta x >= 1\right\} \tag{7}$$

where $\|.\|_2$ denotes the Euclidean vector norm (here $< ., . >$ denotes the dot product). Let Φ have the SVD (5).

Definition 2.1. The ODP δx^o is the solution of the extremal problem

$$J(\delta x) = \|\Phi \delta x\|_2 \to \max_{\delta x} \tag{8}$$

under the constraint (7). One can prove

Lemma 2.1. The optimal perturbation in the sense (8) and (7) is

$$\delta x^o = (+/-)v_1$$

where v_1 is the first right SV of Φ.

2.3.2. Subspaces of ODPs

Introduce

$$\Phi_1 := \Phi - \sigma_1 u_1 v_1^T$$

Consider the optimization problem

$$J_1(\delta x) = \|\Phi_1 \delta x\|_2 \to \max_{\delta x} \tag{9}$$

under the constraint (7). Similar to the proof of Lemma 2.1, one can prove

Lemma 2.2. The optimal perturbation in the sense (9) and (7) is

$$\delta x^o = (+/-)v_2$$

where v_2 is the second right SV of Φ.

By iteration, for

$$\Phi_i := \Phi_{i-1} - \sigma_i u_i v_i^T, i = 1, \ldots, n-1; \Phi_0 = \Phi. \tag{10}$$

applying Lemma 2.2 with slight modifications, one finds that the OPs for $\Phi_i, i = 0, 1, \ldots, n-1$ are $(+/-)v_i, i = 1, 2, \ldots n$..

Theorem 2.1. The optimal perturbation for the matrix $\Phi_{i-1}, i = 1, \ldots, n$ is $(+/-)v_i$ where v_i is the ith leading right SV of $\Phi_{i-1}, i = 1, \ldots, n$.

The OP for Φ_{i-1} will be called the i^{th} OP for Φ (or the i^{th} SOP—singular OP).

Comment 2.2. The OPs, presented above, are optimal in the sense of the Euclidean norm $\|.\|_2$. In practice, there is a need to normalize the state vector (using the inverse of the covariance matrix M). The normalization is done by changing $\delta x' = M^{-1/2}\delta x$, and all the results presented above remain valid s.t. the new $\delta x'$,

$$\|\delta x'\|_2 =< M^{-1/2}\delta x, M^{-1/2}\delta x >=< \delta x, M^{-1}\delta x > \coloneqq \|\delta x\|_{M^{-1}}$$

The weighted norm $\|\delta x\|_{M^{-1}}$ is known as the Mahanalobis norm.

As to the PE, a normalization is also applied in order to have a possibility to compare different variables like density, temperature, velocity … In this situation, the norm for $y = \Phi\delta x$ may be seminorm [5].

2.4. Ensemble forecasting

The idea of ensemble forecasting is that instead of performing "deterministic" forecasts, stochastic forecasts should be made: several model forecasts are performed by introducing perturbations in the filtered estimate or in the models.

Since 1994, NCEP (National Centers for Environmental Prediction, USA) has been running 17 global forecasts per day, with the perturbations obtained using the method of breeding growing perturbations. This ensures that the perturbations contain growing dynamical perturbations. The length of the forecasts allows the generation of outlook for the second week. At the ECMWF, the perturbation method is based on the use of SVs, which grow even faster than the bred or Lyapunov vector perturbations. The ECMWF ensemble contains 50 members [13].

3. Perturbations based on leading EVs and SchVs

3.1. Adaptive filter (AF)

The idea underlying the AF is to construct a filter which uses feedback in the form of the PE signal (innovation) to adjust the free parameters in the gain to optimize the filter performance. If in the KF, the optimality is defined as a minimum mean squared error (MMS), in the AF, optimality is understood in the sense of MMS for the prediction output error (innovation). This definition allows to define the optimality of the filter in the realization space, but not in the probability space as done in the KF.

The optimal gain thus can be determined from solving the optimization problem by adjusting all elements of the filter gain. There are two major difficulties:

i. Instability: As the filter gain is estimated stochastically during the optimization process, the filter may become unstable due to the stochastic character of the filter gain.

ii. Reduction of tuning parameters: For extreme HdS, the number of elements in the filter gain is still very high. Reduction of the number of tunning gain elements is necessary.

3.2. Leading EVs and SchVs as optimal perturbations

Interest on stability of the AF arises soon after the AF has been introduced. The study on the filter stability shows that it is possible to provide a filter stability when the system is *detectable* [8]. For the different parameterized stabilizing gain structures based on a subspace of unstable and neutral EVs, see [8]. As the EVs may be complex and their computation is unstable (Lanczos [14]), in [8], it is proved that one can also ensure a stability of the filter if the space of projection is constructed from a set of unstable and neutral SchVs of the system dynamics. The unstable and neutral real SchVs are referred to as SchVs associated with the unstable and neutral eigenvalues of the system dynamics. The advantage of the real SchVs is that they are real, orthonormal, and their computation is stable. Moreover, the algorithm for estimating dominant SchVs is simple which is based on the power iteration procedure (Sampling-P, see [15]). As to the unstable SVs, although they are real and orthonormal, their computation requires an adjoint operator (the transpose matrix Φ^T). Construction of adjoint code (AC) is a time-consuming and tedious process. Approximating leading SVs without (AC) can be done on the basis of Algorithms 5.2.

3.3. EVs as optimal perturbations in the invariant subspace

Let Φ be diagonalizable. Introduce the set

$$\left\{EV_1(x,\lambda) : x \in C^n, \|x\|_2 = 1, \lambda \in C^1 : \Phi x = \lambda x\right\}. \tag{11}$$

The subspace of $x \in C^n$ satisfying $\Phi x = \lambda x$ for some $\lambda \in C^1$ is known as an invariant subspace of Φ: the matrix Φx acts on to stretch the vector x but conserves the direction of x. Consider the optimization problem

$$\begin{aligned} J(\delta x) = \ & \|\Phi \delta x\|_2 \to \max_{\delta x}, \\ & (\delta x, \lambda) \in EV_1(\delta x, \lambda), \end{aligned} \tag{12}$$

It is seen that the optimal solution is the first EV $x_{ei}(1)$ of Φ with the largest magnitude equal to $|\lambda_1|$. We will call λ_1 a first optimal EV perturbation (denoted as EI-OP).

For a symmetric matrix, the EI-OP coincides with the SOP. The EI-OP is not unique.

By solving the optimization problem (8) s.t.

$$\left\{EV_2(x,\lambda) : x \in C^n, \|x\|_2 = 1 : \Phi x = \lambda x, \lambda \in C^1, |\lambda| < |\lambda_1|\right\}. \tag{13}$$

one finds the second EI-OP $x_{ei}(2)$. In a similar way, by defining

$$\left\{EV_i(x,\lambda) : x \in C^n, \|x\|_2 = 1 : \Phi x = \lambda x, \lambda \in C^1, |\lambda| < |\lambda_{i-1}|\right\}. \tag{14}$$

for $i = 1, 2, .., n-1$, we obtain a sequence of EI-OPs $x_{ei}(i)$, $i = 1, 2, .., n$. The first n_e EI-OPs are unstable SVs.

In general, for a defective case (not diagonalizable), Φ does not have n linearly independent EVs and the independent generalized EVs can serve as "optimal" perturbations to construct a subspace of projection in the AF.

To summarize, let the EV decomposition be

$$X_{ei} J X_{ei}^{-1} = \Phi \tag{15}$$

where J is a matrix of Jordan canonical form, X_{ei}^{-1} is the matrix inverse of X_{ei} (see Golub and Van Loan [9]). The columns of X_{ei} are the EVs of Φ, J is a block diagonal with the diagonal blocks of 1 or 2 dimensions. The rank k decomposition is $X_{ei,1} J_1 \tilde{X}_{ei,1}$ where

$$\begin{aligned} &EV_k \coloneqq X_{ei,1} J_1 \tilde{X}_{ei,1}, \\ &X_{ei} = [X_{ei,1}, X_{ei,2}], J = \text{block diag}[J_1, J_2], \\ &\tilde{X}_{ei} \coloneqq X_{ei}^{-1} = \left[\tilde{X}_{ei,1}^T, \tilde{X}_{ei,2}^T\right], \end{aligned} \tag{16}$$

with $X_{ei,1} \in R^{n \times k}$, $X_{ei,2} \in R^{n \times (n-k)}$. Multiplying the right of EV_k by $X_{ei,1}$ yields $X_{ei,1} J_1$, i.e., we obtain the k largest (in modulus) perturbations in the eigen (invariant) space of Φ. The perturbations being the column vectors of $X_{ei,1}$ (i.e., the k first EVs of Φ) are the first k OPs of Φ in the eigen-invariant subspace (EI-InS).

3.4. Dominant SchVs as OPs in the Schur invariant subspace

The study of Hoang et al. [8] shows that the subspace of projection of the stable filter can be constructed on the basis of all unstable EVs or SchVs of Φ.

Compared to the EVs, the approach based on real Schur decomposition is of preference in practice since the SchVs are real and orthonormal. Moreover, there exists a simple, power iterative algorithm for approaching the set of real leading SchVs. According to Theorem 7.3.1 in Golub and Van Loan [9], the subspace $R[X_{s,1}]$ spanned by the n_u leading SchVs converges to the unique invariant subspace $D_{n_u}(\Phi)$ (called a *dominant* invariant subspace) associated with the eigenvalues $\lambda_1, \ldots, \lambda_{n_u}$ if $|\lambda_{n_u}| > |\lambda_{n_u+1}|$. In this sense, we consider the leading SchVs as OPs (denoted as Sch-OP) in the Schur invariant subspace (Sch-InS).

4. Optimal stochastic perturbation (OSP)

In Section 2, the perturbation δx is deterministic (see Definition 2.1). In practice, it happens that δx is of stochastic nature. For example, the priori information on the FE is an zero mean

http://dx.doi.org/10.5772/intechopen.77273

random vector (RV) with the ECM P. The question arising here is how one determine the OP in such situation and how to find it.

We will consider now δx as an element of the Hilbert space $\mathcal{H}$ of RVs. This space $\mathcal{H}$ is a complete normed linear vector space equipped with the inner product

$$< x, y>_H = E < x, y > \tag{17}$$

where $E(.)$ denotes the mathematical expectation. The norm in $\mathcal{H}$ is defined as

$$\|x\|_H = \sqrt{E < x, x >} \tag{18}$$

All elements of $\mathcal{H}$ are of finite variance and for simplicity, we assume they all have zero mean value.

Introduce the set of RVS

$$S_s(\delta x) \coloneqq \left\{\delta x : \|\delta x\|_H = 1\right\} \tag{19}$$

Definition 4.1. The optimal stochastic perturbation (OSP) δx^o is the solution of the extremal problem

$$J(\delta x) = \|\Phi \delta x\|_H \to \max_{\delta x}, \tag{20}$$

under the constraint (19). One can prove

Lemma 4.1. For $\delta x \in S_s(\delta x)$, there exists $\delta y = \left(\delta y_1, \ldots, \delta y_n\right)^T \in S_s(\delta x)$ such that $\delta x = \sum_{k=1}^n v_k \delta y_k$.

Lemma 4.2. The optimal perturbation in the sense (20) and (19) is

$$\delta x^o = \psi v_1$$

where ψ is a RV with zero mean and unit variance, v_1 is the first right SV of Φ.

Comment 4.1. Comparing the ODP with OSP shows that if the ODS is the first right SV (defined up to the sign), the OSP is an ensemble of vectors lying in the subspace of the first right SV with the lengths being the samples of the RV of zero mean and unit variance.

Introduce

$$\Phi_1 \coloneqq \Phi - \sigma_1 u_1 v_1^T$$

and consider the objective function

$$J(\delta x) = \|\Phi_1 \delta x\|_H \to \max_{\delta x}, \tag{21}$$

Lemma 4.3. The optimal perturbation in the sense (21) and (19) is $\delta x^o = \psi v_2$, where ψ is an RV with zero mean and unit variance, v_2 is the first right SV of Φ.

By iteration, for

$$\Phi_i := \Phi_{i-1} - \sigma_i u_i v_i^T, i = 1, \ldots, n-1; \Phi_0 = \Phi. \tag{22}$$

applying Lemma 4.3 with slight modifications, one finds that the OSP for $\Phi_k, k = 0, 1, \ldots, n-1$ are $\psi v_k, k = 1, 2, \ldots n.$, where ψ is an RV with zero mean and unit variance, v_k is the k^{th} right SV of Φ.

Theorem 4.1. The optimal perturbation for the matrix $\Phi_{i-1}, i = 1, \ldots, n$ is ψv_i, where ψ is an RV with zero mean and unit variance, v_k is the k^{th} right SV of Φ.

5. Simultaneous stochastic perturbations (SSP)

In [16], Spall proposes a simultaneous perturbation stochastic approximation (SPSA) algorithm for finding optimal unknown parameters by minimizing some objective function. The main feature of the simultaneous perturbation gradient approximation (SPGA) resides in the way to approximate the gradient vector (in average): a sample gradient vector is estimated by perturbing simultaneously all components of the unknown vector in a stochastic way. This method requires only two or three measurements of the objective function, regardless of the dimension of the vector of unknown parameters. In a study by Hoang and Baraille [17], the idea of the SPGA is described in detail, with a wide variety of applications in engineering domains. The application to estimation of ECM in the filtering problem is given in the work done by Hoang and Baraille [18]. In the research by Hoang and Baraille [19], a simple algorithm for estimating the elements of an unknown matrix as well as the way to decompose the estimated matrix into a product of two matrices, under the condition that only the matrix-vector product is accessible, has been proposed.

5.1. Theoretical background of SPGA: gradient approximation

The component-wise perturbation is a method for numerical computation of the cost function with respect to the vector of unknown parameters. It is based on the idea to perturb separately each component of the vector of parameters. For very HdS, this technique is impossible to implement. An alternative to the component-wise perturbation is the SSP approach.

Let

$$DJ(\theta_0) = \left[\frac{\partial J(\theta_0)}{\partial \theta_1}, \ldots, \frac{\partial J(\theta_0)}{\partial \theta_{n_\theta}}\right]^T$$

denote the gradient of $J(\theta)$ computed at $\theta = \theta_0$. Suppose $\Delta_j, j = 1, \ldots, n$ are RVs independent identically distributed (i.i.d) according to the Bernoulli law which assumes two values +1 or -1 with equal probabilities 1/2. It implies that

$$E(\Delta_j) = 0, E(\Delta_j)^2 = 1, E\left(\Delta_j^{-1}\right) = 0, E\left(\Delta_j^{-1}\right)^2 = 1, j = 1, 2, \ldots, n \tag{23}$$

Suppose $J(\theta)$ is infinitely differentiable at $\theta = \theta_0$. Using a Taylor series expansion,

$$\Delta J \coloneqq J(\theta_0 + \overline{\delta}\theta) - J(\theta_0) = \overline{\delta}\theta^T DJ(\theta_0) + (1/2)\overline{\delta}\theta^T D^2 J(\theta_0)\overline{\delta}\theta + \ldots \tag{24}$$

where $D^2 J(\theta_0)$ is the Hessian matrix computed at $\theta \coloneqq \theta_0$. For the choice

$$\overline{\delta}\theta \coloneqq (\delta\theta_1, \ldots, \delta\theta_{n_\theta})^T = c\overline{\Delta}, \overline{\Delta} = (\Delta_1, \Delta_2, \ldots, \Delta_{n_\theta})^T, \tag{25}$$

c is a small positive value, from Eq. (24)

$$\Delta J(\theta_0) = c\overline{\Delta}^T DJ(\theta_0) + (c^2/2)\overline{\Delta}^T D^2 J(\theta_0)\overline{\Delta} + \ldots$$

Dividing both sides of the last equality by $\delta\theta_k = c\Delta_k$ implies

$$\begin{aligned} \Delta J(\theta_0)/\delta\theta_k = DJ(\theta_0)^T\overline{\Delta}^k + (c/2)\overline{\Delta}^{k,T} D^2 J(\theta_0)\overline{\Delta} + \ldots \\ \overline{\Delta}^k \coloneqq \left(\Delta_1\Delta_k^{-1}, \ldots, 1, \ldots, \Delta_n\Delta_k^{-1}\right)^T \end{aligned} \tag{26}$$

Taking the mathematical expectation for both sides of the last equation yields

$$\begin{aligned} E\left[\Delta J(\theta_0)\delta\theta_k^{-1}\right] = \\ DJ(\theta_0)^T E\left(\overline{\Delta}^k\right) + (c/2)E\left[\overline{\Delta}^T D^2 J(\theta_0)\overline{\Delta}^k\right] + \ldots \end{aligned} \tag{27}$$

One sees that from the assumptions on $\overline{\Delta}$, $E\left(\overline{\Delta}^k\right) = (0, \ldots, 1, \ldots, 0)^T$ it follows $DJ(u_0)^T E\left(\overline{\Delta}^k\right) = \partial J(\theta_0)/\partial\theta_k$. Moreover, as all the moments of the Bernoulli variables Δ_i and Δ_i^{-1} are finite, $E\left[\overline{\Delta}^T D^2 J(\theta_0)\overline{\Delta}^k\right] = 0$ since there exists a finite $D^2 J(\theta_0)$, one concludes that

$$E\left[\Delta J(\theta_0)\delta\theta_k^{-1}\right] = \partial J(\theta_0)/\partial\theta_k + O(c^2) \tag{28}$$

The result expressed by Eq. (28) constitutes a basis for approximating the gradient vector by simultaneous perturbation. The left of Eq. (28) can be easily approximated by noticing that for an ensemble of L i.i.d samples $\left[\overline{\Delta}^{(1)}, \ldots, \overline{\Delta}^{(L)}\right]$, we can generate the corresponding ensemble of L i.i.d sample estimates for the gradient vector at $\theta = \theta_0$,

$$DJ^{(l)}(\theta_0) = \left[\frac{\Delta J^{(l)}(\theta_0)}{c\Delta_1^{(l)}}, \ldots, \frac{\Delta J^{(l)}(\theta_0)}{c\Delta_n^{(l)}}\right]^T, l = 1, 2, \ldots, L \tag{29}$$

where $\Delta_k^{(l)}$ is the k^{th} component of the l^{th} sample $\overline{\Delta}^{(l)}$. The left of Eq. (28) is then well approximated by averaging L sample gradients in Eq. (29),

$$E[\Delta J(\theta_0)/\delta\theta_k] \approx (1/L)\sum_{l=1}^{L}\eta_k^{(l)},\ \eta_k^{(l)} := \frac{\Delta J^{(l)}(\theta_0)}{c\Delta_k^{(l)}}, \tag{30}$$

Introduce the notations

$$\overline{m}_k := E[\Delta J(\theta_0)/\delta\theta_k],\ m_k(L) := (1/L)\sum_{l=1}^{L}\eta_k^{(l)},\ e_k := m_k(L) - \overline{m}_k. \tag{31}$$

Theorem 1 (Hoang and Baraille [19]) states that the estimate $m(L) := (m_1(L), \ldots, m_{n_\theta}(L))^T$ converges to the gradient vector $DJ(\theta_0)$ as $L \to \infty$ and $c \to 0$ with the order $O(1/L)$ where

$$m(L) := (1/L)\sum\nolimits_{l=1}^{L}\eta^{(l)}.$$

5.2. Algorithm for estimation of an unknown matrix

Let $\overline{\Delta} := (\Delta_1, \ldots, \Delta_n)^T$, Δ_i, $i = 1, \ldots, n$ be Bernoulli independent and identically distributed (i.i.d.) variables assuming two values ± 1 with equal probability 1/2. Introduce $\left[\overline{\Delta}\right]^{-1} := (1/\Delta_1, \ldots, 1/\Delta_n)^T$, $\overline{\Delta}_c := c\overline{\Delta}$, $c > 0$ is a small positive value.

Algorithm 5.1. Suppose it is possible to compute the product $\Phi x = b(x)$ for a given x. At the beginning let $l = 1$. Let the value u be assigned to the vector x, i.e., $x := u$, L be a (large) fixed integer number.

Step 1. Generate $\overline{\Delta}^{(l)}$ whose components are l^{th} samples of the Bernoulli i.i.d. variables assuming two values +/− 1 with equal probabilities 1/2;

Step 2. Compute $\delta b^{(l)} = \Phi\left(u + \overline{\Delta}_c^{(l)}\right) - \Phi u$, $\overline{\Delta}_c^{(l)} = c\overline{\Delta}^{(l)}$.

Step 3. Compute $g_i^{(l)} = \delta b_i^{(l)}\left[\overline{\Delta}_c^{(l)}\right]^{-1}$, δb_i is the i^{th} component of δb, $g_i^{(l)}$ is the column vector consisting of derivative of $b_i(u)$ w.r.t. u, $i = 1, \ldots, m$.

Step 4. Go to Step 1 if $l < L$. Otherwise, go to Step 5.

Step 5. Compute

$\widehat{g}_i = \frac{1}{L}\sum_{l=1}^{L} g_i^{(l)}$, $i = 1, \ldots, m$, $\widehat{\Phi}(L) := D_x b = \left[\widehat{g}_1, \ldots, \widehat{g}_m\right]^T$.

5.3. Operations with Φ and its transpose

Algorithm 5.1 allows to store $\widehat{\Phi}(L)$ as composed from the two ensembles of vectors elements:

$$\begin{aligned} En_L(\delta x) &:= \left[\delta x^{(1)}, \ldots, \delta x^{(L)}\right],\ \delta x^{(l)} = c\overline{\Delta}^{(l)}, \\ En_L(\delta b) &:= \left[\delta b^{(1)}, \ldots, \delta b^{(L)}\right],\ \delta b^{(l)} = \left(\delta b_1^{(l)}, \ldots, \delta b_m^{(l)}\right)^T,\ l = 1, \ldots, L. \end{aligned} \tag{32}$$

The product $z = \widehat{\Phi}(L)y$, $y \in R^n$, can be performed as $z_i = \sum_{k=1}^{n} \widehat{\phi}_{ik} y_k = \sum_{k=1}^{n} \left[\frac{1}{L}\sum_{l=1}^{L} \frac{\delta b_i^{(l)}}{\delta x_k^{(l)}}\right] y_k$, or in a more compact form

$$z = \frac{1}{L}\sum_{l=1}^{L} \alpha_l \delta b^{(l)}, \alpha_l := \sum_{k=1}^{n} \frac{y_k}{\delta x_k^{(l)}}. \tag{33}$$

Eq. (33) allows to perform $z = \widehat{\Phi}(L)y$ with $L(m+2n)+1$ elementary operations.

Similarly, computation of z_i of $z = \widehat{\Phi}^T(L)y$, $y \in R^m$ is performed as

$$z_i = \frac{1}{L}\sum_{l=1}^{L} \frac{1}{\delta x_i^{(l)}} \sum_{k=1}^{m} \delta b_k^{(l)} y_k, i = 1, \ldots, n \tag{34}$$

5.4. Estimation of decomposition of Φ

Let Φ be a matrix of dimensions $(m \times n)$. For definiteness, let $m \le n$ with $\text{rank}(\Phi) = m$. We want to find the best approximation for Φ among members of the class of matrices

$$\Phi_e = AB^T, A \in R^{m \times r}, B \in R^{n \times r}. \tag{35}$$

under the constraint

Condition (C) A, B are matrices of dimension $m \times r, r \times n, r \le m$, $\text{rank}\left(AB^T\right) = r$.

Under the condition (C), the optimization problem is formulated as

$$J(A,B) = \|\Phi - \Phi_e\|_F^2 = \left\|\Phi - AB^T\right\|_F^2 \to \min_{(A,B)}, \tag{36}$$

where $\|.\|_F$ denotes the Frobenius matrix norm. Consider Φ and let $U\Sigma V^T$ be SVD of Φ (5). Let $\tilde{\Phi} = \Phi + \Delta\Phi, \tilde{\Phi} = \tilde{U}\tilde{\Sigma}\tilde{V}^T$ and $\tilde{\sigma}_1 \ge \tilde{\sigma}_2 \ldots \ge \tilde{\sigma}_m$, $\tilde{\sigma}_k$ be the k^{th} singular value of $\tilde{\Phi}$. Then, we have

$$J(A_o, B_o) = \sum_{k=r+1}^{m} \sigma_k^2 \tag{37}$$

where $A_o B_o^T$ is a solution to the problem (36) s.t Condition (C) (Theorem 3.1 of Hoang and Baraille [19]).

Theorem 3.1. Hoang and Baraille [19] implies that $\Phi_e^o := A_o B_o^T$ is equal to the matrix formed by truncating the SVD of Φ to its first r SVs and singular values. It allows to avoid storing elements of the estimate $\widehat{\Phi}(L)$ of Φ (their number is of order $O(10^{m \times n})$).

5.4.1. Decomposition algorithms

Let the elements of Φ (or $\widehat{\Phi}$) be available (may be in algorithmic form). By perturbing stochastically simultaneously all the elements of A and B, one can write out the iterative algorithm for estimating the elements of A and B. For more detail, see Hoang and Baraille [19].

5.4.2. Iterative decomposition algorithm

Another way to decompose the matrix Φ is to solve iteratively the following optimization problems

Algorithm 5.2

At the beginning let $i = 1$.

Step 1. For $i = 1$, solve the minimization problem

$$J_1 = \left\|\Phi^1 - ab^T\right\|_F^2 \to \min_{a,b}, \quad a \in R^m, b \in R^n.$$
$$\Phi^1 := \Phi, \quad \operatorname{rank}\left(ab^T\right) = 1$$

Its solution is denoted as $\widehat{a}(i), \widehat{b}(i)$.

Step 2. For $i < r$, put $i := i + 1$ and solve the problem

$$J_{i+1} = \left\|\Phi^i - ab^T\right\|_F^2 \to \min_{a,b},\ a \in R^m, b \in R^n.\ \Phi^i := \Phi - \sum_{k=1}^{i-1} \widehat{a}(k)\widehat{b}^T(k),$$
$$\operatorname{rank}\left(ab^T\right) = 1$$

Step 3. If $i = r$, compute

$$\widehat{\Phi} = \widehat{A}(r)\widehat{B}^T(r),\ \widehat{A}(r) = [\widehat{a}(1), \ldots, \widehat{a}(r)],\ \widehat{B}(r) = \left[\widehat{b}(1), \ldots, \widehat{b}(r)\right].$$

and stop. Otherwise, go to *Step 2*.

From Theorem 3.2 of Hoang and Baraille [19], the couple $\widehat{A}(r), \widehat{B}(r)$ is a solution for the problem (36)(C).

6. Numerical example

Consider the matrix $\Phi \in R^{2\times 2}$

$$\phi_{11} = 5, \phi_{12} = 7, \phi_{21} = -2, \phi_{22} = -4.$$

The singular values and the right SVs for Φ are displayed in **Table 1** which are obtained by solving the classical equations for eigenvalues of $\Phi^T\Phi$.

First, we apply Algorithm 5.1 to estimate the matrix Φ. **Figure 1** shows the estimates produced by Algorithm 5.1. It is seen that the estimates are converging quickly to the true elements of Φ.

Next Algorithm 5.2 (Iterative Decomposition Algorithm) has been applied to estimate the decomposition of the matrix Φ. After each i^{th} iteration, the algorithm yieds $\widehat{b}^{(i)}, \widehat{c}^{(i)}$.

The different OPs are shown in **Table 2** where $x_{sv}^r(i)$, $x_{ei}(i)$, and $x_{sch}(i)$ are the theoretical SV-POs, EI-POs, Sch-POs. The vectors $\widehat{x}_{sch}(i)$ are the components of X_t computed by *Algorithm* 3.1

Element	ϕ_{ij}	$\widehat{\phi}_{ij}$	$\widehat{\phi}_{ij^{(1)}}$	$\widehat{\phi}_{ij^{(2)}}$
(1,1)	5.00	4.677	4.914	−0.237
(1,2)	7.00	7.07	7.662	−0.592
(2,1)	−2.00	−2.041	−1.08	−0.962
(2,2)	−4.00	−4.081	−1.683	−2.398

Table 1. Estimates of Φ obtained by Algorithm 5.2.

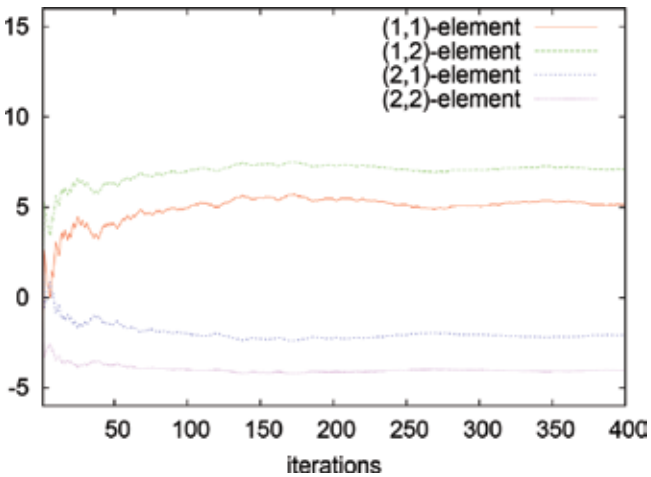

Figure 1. Estimates of elements of the matrix Φ.

Perturbations	Vector	Predictor	Amplification
$x^r_{sv}(1)$	$(0.554, 0.833)^T$	$(8.597, -4.438)^T$	9.676
$x^r_{sv}(2)$	$(0.833, -0.554)^T$	$(0.284, 0.551)^T$	0.62
$x_{ei}(1)$	$(0.962, -0.275)^T$	$(2.885, -0.824)^T$	3
$x_{ei}(2)$	$(0.707, -0.707)^T$	$(-1.414, 1.414)^T$	2
$x_{sch}(1)$	$(0.707, 0.707)^T$	$(8.485, -4.243)^T$	9.487
$x_{sch}(2)$	$(0.707, -0.707)^T$	$(-1.414, 1.414)^T$	2
$\widehat{x}_{sch}(1)$	$(0.962, -0.275)^T$	$(-8.1, 4.4)^T$	9.22
$\widehat{x}_{sch}(2)$	$(-0.275, -0.962)^T$	$(2.885, -0.824)^T$	3
$\widehat{c}^{(1)}_n$	$(0.54, 0.842)^T$	$(8.592, -4.447)^T$	9.674
$\widehat{c}^{(2)}_n$	$(0.372, 0.928)^T$	$(8.358, -4.457)^T$	9.472

Table 2. Different OPs.

(Sampling-P). As to $\widehat{c}_n^{(i)}$, they are the results of normalization (with the unit Euclidean norm) of $\widehat{c}^{(i)}$.

Looking at the first OPs, one sees that $x_{sv}^r(1)$, $x_{sch}(1)$, $\widehat{x}_{sch}(1)$, and $\widehat{c}_n^{(1)}$ produce almost the same amplification. The first $x_{ei}(1)$ has the amplification three times less than those of $x_{sv}^r(1)$, $x_{sch}(1)$. The second $x_{sch}(2)$ is much less opimal than $x_{sv}^r(2)$ and $\widehat{c}_n^{(2)}$. By comparing $x_{sv}^r(i)$ with $\widehat{c}_n^{(i)}$ for $i = 1, 2$, one concludes that the obtained results justify the correctness of Theorem 3.1 of Hoang and Baraille [19]. Mention that only $\widehat{x}_{sch}(i)$ and $\widehat{c}_n^{(i)}$ can be calculated for HdS.

In **Table 1**, we show the results obtained by Algorithm 5.2 after two consecutive iterations (matrix estimation in R^1 subspace). The elements of the true $\Phi = \left[\phi_{ij}\right]$ are displayed in the second column, whereas their estimates—in the third column,

$$\widehat{\Phi} = \widehat{\Phi}^{(1)} + \widehat{\Phi}^{(2)} = \sum_{i=1}^{2} \widehat{b}^{(i)}\widehat{c}^{(i),T}\widehat{\Phi}^{(i)} := \left[\widehat{\phi}_{ij}^{(i)}\right] = \widehat{b}^{(i)}\widehat{c}^{(i),T}$$

The estimates, resulting from the first iteration, are the elements of $\widehat{\Phi}^{(1)}$ (**Table 1**, column 4). After the first iteration, $\Phi^{(2)} := \Phi - \widehat{\Phi}^{(1)} = b^{(2)}c^{(2),T}$ and the optimization yields the estimates $\widehat{\phi}_{ij}^{(2)}$ displayed in the column 5. From the columns 4–5, one sees that the first iteration allows to well estimate the two biggest elements $\Phi_{11} = 5$, $\Phi_{12} = 7$. In the similar way, the second iteration captures the two biggest elements of $\Phi^{(2)}$.

7. Assimilation in high-dimensional ocean model MICOM

7.1. Ocean model MICOM

To see the impact of optimal SchVs in the design of filtering algorithm for HdS, in this section, we present the results of the experiment on the Hd ocean model MICOM (Miami Isopycnal Ocean Model). This numerical experiment is identical to that described in Hoang and Baraille [15]. The model configuration is a domain situated in the North Atlantic from 30°N to 60°N and 80°W to 44°W; for the exact model domain and some main features of the ocean current produced by the model, see and Baraille [15]. The system state $x = (h, u, v)$ where $h = h(i, j, k)$ is a layer thickness and $u = u(i, j, k), v = v(i, j, k)$ are two velocity components. Mention that after discretization, the dimension of the system state is $n = 302400$. The observations available at each assimilation instant are the sea surface height (SSH) with dimension $p = 221$.

7.1.1. Data matrix based on dominant Sch-Ops

The filter is a reduced-order filter (ROF) with the variable h as a reduced state and u, v are calculated on the basis of the geostrophy hypothesis. To obtain the gain in the ROF, first the *Algorithm 3.1* has been implemented to generate an ensemble of dominant SchVs (totally 72 SchVs, denoted as $En(SCH)$). The sample ECM $M^d(SCH)$ is computed on the basis of the

$En(SCH)$. Due to rank deficiency, the sample $M^d(SCH)$ is considered only as a data matrix. The optimization procedure is applied to minimize the distance between the data matrix $M^d(SCH)$ and the structured parametrized ECM $M(SCH) = M_v \otimes M_h$ which is written in the form of the Schur product of two matrices $M(SCH) = M_v(\theta) \otimes M_h(\theta)$. Here, M_v is the vertical ECM, M_h is the horizontal ECM [18]), (θ) is a vector of unknown parameters. Mention that the hypothesis on separability of the vertical and horizontal variables in the ECM is not new in the meteorology [20]. The gain is computed according to Eq. (3) with $R = \alpha I$, $\alpha > 0$ is a small positive value. The ROF is denoted as PEF (SSP).

7.1.2. Data matrix based on SSP approach

The second data matrix $M^d(SSP)$ is obtained by perturbing the system state according to the SSP method. The SSP samples are simulated in the way similar to that described above for generating $En(SCH)$, with the difference that the perturbation components $\delta h^{(l)}(i,j,k)$ are the i. i.d. random Bernoulli variables assuming two values ± 1 with the equal probability 1/2. The same optimization procedure has been applied to estimate $M(SSP)$. The obtained ROF is denoted as PEF (SSP).

Figure 2 shows the evolution of estimates for the gain coefficients $k(1)$ computed from the estimated coefficients of $\widehat{c}_{kl}$ of $M^d(SCH)$ and $M^d(SSP)$ on the basis of $En(SCH)$ (curve" schur") and $En(SSP)$ (curve "random"), during model integration. It is seen that two coefficients are evolved in nearly the same manner, of nearly the same magnitude as that of $k(1)$ in the CHF (Cooper-Haines filter, Cooper and Haines [21]). Mention that the CHF is a filter widely used in the oceanic data assimilation, which projects the PE of the surface height data by lifting or lowering of water columns.

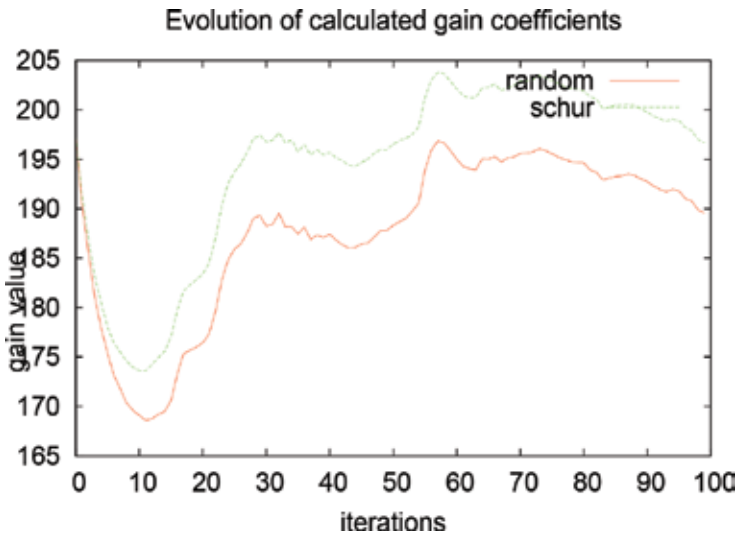

Figure 2. Evolution of estimates for the gain coefficients $k(1)$ computed from $\widehat{c}_{kl}$ on the basis of $En(SCH)$ (curve "Schur") and $En(SSP)$ (curve "random"), during model integration. It is seen that two coefficients are evolved in nearly the same manner, of nearly the same magnitude as that of $k(1)$ in the CHF. The same picture is obtained for other $c_k, k = 2, 3, 4$.

The same pictures are obtained for the estimates $\widehat{c}_k, k = 2, 3, 4$. Mention that in the CHF, $c_2 = c_3 = 0$.

7.2. Performance of different filters

In **Table 3**, the performances of the three filters are displayed. The errors are the averaged (spatially and temporally) rms of PE for the SSH and for the two velocity components u and v.

The results in **Table 3** show that two filters PEF (SCH) and PEF (SSP) are practically of the same performance, and their estimates are much better compared to those of the CHF, with a slightly better performance for the PEF (SSP). We note that as the PEF (SCH) is constructed on the basis of an ensemble of samples tending to the first dominant SchV, its performance must be theoretically better than that of the PEF (SSP). The slightly better performance of PEF (SSP) (compared to that of PEF (SCH)) may be explained by the fact that the best theoretical performance of PEF (SCH) can be obtained only if the model is linear, stationary, and the number of PE samples in $En(SCH)$ at each iteration must be large enough. The ensemble size of $En(SCH)$ in the present experiment is too small compared with the dimension of the MICOM model.

rms	CHF (cm)	PEF (SCH) (cm/s)	PEF (RAN) (cm/s)
ssh(fcst)	7.39	5.09	4.95
u(fcst)	7.59	5.36	5.29
v(fcst)	7.72	5.73	5.64

Table 3. rms of PE for *ssh*, and u, v velocity components.

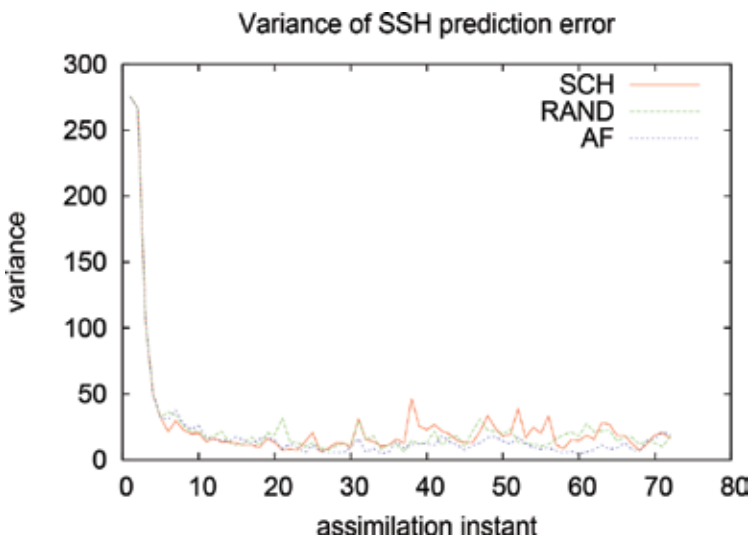

Figure 3. Variance of PE resulting from three filters PEF (SCH), PEF (SSP), and AF. It is seen that the AF yields much better performance compared to the two other nonadaptive filters PEF (SCH) and PEF (SSP).

To illustrate the efficiency of adaptation, in **Figure 3**, we show the cost functions (variances of innovation) resulting from the three filters PEF (SCH), PEF (SSP), and AF (i.e., APEF based on PEF (SCH); the same performance is observed for the AF based on PEF (SSP)). Undoubtedly, the adaptation allows to improve considerably the performances of nonadaptive filters.

8. Conclusion remarks

We have presented in this chapter the different types of OPs—deterministic, stochastic, or optimal, the invariant subspaces of the system dynamics. The ODPs and OSPs play an important role in the study on the predictability of the system dynamics as well as in construction of optimal OFS for environmental geophysical systems.

One other class of perturbation known as SSP is found to be a very efficient tool for solving optimization and estimation problems, especially with Hd matrices and in computing the optimal perturbations.

The numerical experiments presented in this chapter confirm the important role of the different types of OPs in the numerical study of Hd assimilation systems.

Author details

Hong Son Hoang* and Remy Baraille

*Address all correspondence to: hhoang@shom.fr

SHOM/HOM/REC, Toulouse, France

References

[1] Myhrvold N. Moore's Law Corollary: Pixel Power. New York Times. June 7, 2006 [Retrieved: 2011-11-27]

[2] Kalman RE. A new approach to linear filtering and prediction problems. Transactions of the ASME-Journal of Basic Engineering. 1960;**82**:35-45

[3] Lorenz EN. Deterministic non-periodic flow. Journal of the Atmospheric Sciences. 1963;**20**: 130-141

[4] Lorenz EN. Atmospheric predictability experiments with a large numerical model. Tellus. 1982;**34**:505-513

[5] Palmer TN, Barkmeijer J, Buizza R, Petroliagis T. The ECMWF ensemble prediction system. Quarterly Journal of the Royal Meteorological Society. 1997;**4**(4):301-304

[6] Toth Z, Kalnay E. Ensemble forecasting at NMC: The generation of perturbations. Bulletin of the American Meteorological Society. 1993;**74**:2317-2330

[7] Hoang HS, De Mey P, Talagrand O, Baraille R. A new reduced-order adaptive filter for state estimation in high dimensional systems. Automatica. 1997;**33**:1475-1498

[8] Hoang HS, Talagrand O, Baraille R. On the design of a stable adaptive filter for high dimensional systems. Automatica. 2001;**37**:341-359

[9] Golub GH, Van Loan CF. Matrix Computations. 2nd ed. Baltimore-London: Johns Hopkin Press, 1993

[10] Hoang HS, Baraille R, Talagrand O. On the stability of a reduced-order filter based on dominant singular value decomposition of the systems dynamics. Automatica. 2009; **45**(10):2400-2405. DOI: 10.1016/j.automatica.2009.06.032

[11] Doucet A, Johansen AM. A tutorial on particle filtering and smoothing: fifteen years later (PDF). Technical report. Department of Statistics, University of British Columbia; 2008

[12] Evensen G. The ensemble Kalman filter: Theoretical formulation and practical implementation. Ocean Dynamics. 2003;**53**:343-367

[13] Kalnay E. Atmospheric Modeling, Data Assimilation and Predictability. Cambridge University Press; 2002

[14] Lanczos C. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. Journal of Research of the National Bureau of Standards. 1950;**45**:255-282

[15] Hoang HS, Baraille R. Prediction error sampling procedure based on dominant Schur decomposition. Application to state estimation in high dimensional oceanic model. Applied Mathematics and Computation. 2011;**218**(7):3689-3709. DOI: 10.1016/j.amc.2011.09.012

[16] Spall JC. Accelerated second-order stochastic optimization using only function measurements. In: Proceedings of 36th IEEE Conference on Decision and Control; 1997. pp. 1417-1424

[17] Hoang HS, Baraille R. Stochastic simultaneous perturbation as powerful method for state and parameter estimation in high dimensional systems. In: Baswell AR, editor. Advances in Mathematics Research. Vol. 20. New-York: Nova Science Publishers; 2015. pp. 117-148

[18] Hoang HS, Baraille R. On the efficient low cost procedure for estimation of high-dimensional prediction error covariance matrices. Automatica. 2017;**83**:317-330

[19] Hoang HS, Baraille R. A simple numerical method based simultaneous stochastic perturbation for estimation of high dimensional matrices. Multidimensional Systems and Signal Processing. 2018. DOI: 10.1007/s11045-018-0551-y

[20] Daley R. The effect of serially correlated observation and model error on atmospheric data assimilation. Monthly Weather Review. 1992;**120**:165-177

[21] Cooper M, Haines K. Altimetric assimilation with water property conservation. Journal of Geophysical Research. 1996;**101**:1059-1077

Chapter 5

Periodic Perturbations: Parametric Systems

Albert Morozov

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.79513

Abstract

We are not going to present the classical results on linear parametric systems, since they are widely discussed in literature. Instead, we shall consider nonlinear parametric systems and discuss the conditions of new motion existence in the resonance zones: the regular ones (on an invariant torus) and the irregular ones (on a quasi-attractor). On the basis of the self-oscillatory shortened system which determines the topology of resonance zones, we study the transition from a resonance to a non-resonance case under a change of the detuning. We then apply our results to some concrete examples. It is interesting to study the behavior of a parametric system when the ring-like resonance zone is contracted into a point, i.e., to describe the bifurcations which occur in the course of transition from the plain nonlinear resonance to the parametric one. We are based on article, and we follow a material from the book.

Keywords: resonances, quasi-attractor, periodic solves, parametric perturbations

1. Introduction

Consider the following system:

$$\begin{aligned} \frac{dx}{dt} &= \frac{\partial H(x,y)}{\partial y} + \varepsilon g(x,y,\nu t), \\ \frac{dy}{dt} &= -\frac{\partial H(x,y)}{\partial x} + \varepsilon f(x,y,\nu t), \end{aligned} \tag{1}$$

where $\varepsilon > 0$ is a small parameter, ν is perturbation frequency, and g, f are continuous periodic functions of period 2π with respect to $\varphi = \nu t$. The Hamiltonian H as well as f and g will be assumed to be sufficiently smooth in a domain $G \subset R^2 \times S^1$ (or $G \subset R^1 \times S^1 \times S^1 = R^1 \times T^2$).

Also, we shall assume that the unperturbed ($\varepsilon = 0$) Hamiltonian system is nonlinear and has at least one cell D filled with closed phase curves.

We especially emphasize the following condition.

Condition A. $\frac{\partial g}{\partial x} + \frac{\partial f}{\partial y} \not\equiv 0$.

This implies that system (1) is nonconservative.

Along with (1), we shall consider the autonomous system:

$$\begin{aligned} \frac{dx}{dt} &= \frac{\partial H(x,y)}{\partial y} + \varepsilon g_0(x,y) \\ \frac{dy}{dt} &= -\frac{\partial H(x,y)}{\partial x} + \varepsilon f_0(x,y), \end{aligned} \tag{2}$$

where $g_0 = < g >_\varphi$ and $f_0 = < f >_\varphi$.

We also assume the following condition.

Condition B. System (2) has a finite set of rough limit cycles (LCs) in cell D.

Changing the variables x, y to the action I and angle θ, we obtain the system in the form

$$\begin{aligned} \dot{I} &= \varepsilon F_1(I,\theta,\varphi) \\ \dot{\theta} &= \omega(I) + \varepsilon F_2(I,\theta,\varphi) \\ \dot{\varphi} &= \nu, \end{aligned} \tag{3}$$

where

$$F_1 \equiv f x'_\theta - g y'_\theta, \quad F_2 \equiv -f x'_I + g y'_I \tag{4}$$

are periodic of period 2π with respect to θ and φ. System (3) is defined on the direct product $\Delta \times S^1 \times S^1 = \Delta \times T^2$, where T^2 is two-dimensional torus, $\Delta = \left(I^-, I^+\right), \quad I^\pm = I\left(h^\pm\right)$.

The definition of resonance. We say that in system (3) a resonance takes place if

$$\omega\left(I_{pq}\right) = (q/p)\nu, \tag{5}$$

where p, q are relatively prime integer numbers.

The energy level $I = I_{pq}$ ($H(x,y) = h_{pq}$) of the unperturbed system is called the resonance.

The behavior of solutions in the neighborhoods

$$U_\mu = \left\{(I,\theta) : I_{pq} - C\mu < I < I_{pq} + C\mu, \quad 0 \le \theta \le 2\pi, C = const\right\}, \quad \mu = \sqrt{\varepsilon}$$

of individual resonance levels $I = I_{pq}$ $\left(H(x,y) = h_{pq}\right)$ can be derived, up to the terms $O\left(\mu^2\right)$, from the pendulum-type equation [1, 2]

$$\frac{d^2v}{d\tau^2} - bA_0(v; I_{pq}) = \mu\sigma(v; I_{pq})\frac{dv}{d\tau}, \tag{6}$$

$$\begin{aligned} b &= d\omega(I_{pq})/dI, \tau = \mu t, A_0(v; I_{pq}) = \frac{1}{2\pi p}\int_0^{2\pi p} F(I_{pq}, v + q\varphi/p, \varphi)d\varphi, \\ \sigma(v, I_{pq}) &= \frac{1}{2\pi p}\int_0^{2\pi}\left(\frac{\partial g(x,y,\varphi)}{\partial x} + \frac{\partial f(x,y,\varphi)}{\partial y}\right)d\varphi, \end{aligned} \tag{7}$$

where $X = X(I_{pq}, v + q\varphi/p), \quad Y = Y(I_{pq}, v + q\varphi/p)$ is the unperturbed solution on the level $I = I_{pq}$. For nondegenerate resonance zones we consider here, it holds that $b \neq 0$. Functions $A_0(v; I_{pq}), \sigma(v; I_{pq})$ are periodic of period $2\pi/p$ with respect to v.

From Eq. (7) follows.

Theorem 1

If the divergence of the vector field of Eq. (6) depends on v, then the divergence of the vector field of the original system (1) contains terms which depend on both the time t and the spatial coordinates.

In many cases the converse is also true. For example, it holds for the system

$$dx/dt = y, \;\; dy/dt = -x - x^3 + \left(P_1 + P_2x^2 + P_3x\sin(\nu t)\right)y + P_4\sin(\nu t). \tag{8}$$

The terms mentioned in Theorem 1 are called nonlinear parametric terms. Our goal is to study systems of the form (1) with such terms. The existence of those leads to new motions in resonance zones [1–3]. We shall demonstrate these motions on examples.

2. Investigation of Eq. (6)

The following representations hold:

$$\begin{aligned} A_0(v; I_{pq}) &= A_*(v; I_{pq}) + B(I_{pq}), \quad B = < A_0 >_v, \\ \sigma(v; I_{pq}) &= \sigma_*(v; I_{pq}) + B_1(I_{pq}), \quad B_1 = < \sigma >_v, \end{aligned} \tag{9}$$

where $B(I)$ is the generating function of the autonomous system (2) and $B_1(I)$ is the derivative of $B(I)$. We shall focus on the case when σ is sign-alternating. In this case, from Eq. (9) follows the inequality:

$$|B_1(I_{pq})| < \max_v|\sigma_*(v, I_{pq})|. \tag{10}$$

When studying the pendulum Eq. (6), we shall distinguish two cases: (I) $B(I_{pq}) \neq 0$ and (II) $B(I_{pq}) = 0$.

In case II system (2) has a rough limit cycle (LC) in a neighborhood of the level $H(x, y) = h_{pq}$. There is no such cycle in case I.

Case I. Neglecting terms of order μ in Eq. (6), we arrive at the integrable equation

$$d^2v/d\tau^2 - bA_0(v, I_{pq}) = 0 \tag{11}$$

If $|B(I_{pq})| > max_v|A_*(v, I_{pq})|$, then Eq. (11) has no equilibrium states. The resonance level $I = I_{pq}$ is then referred to as passable. Note that the term "passable" has its origin in the topology of the resonance zone, as opposed to the same term used in physics, where "passing" stands for a change in perturbation frequency ν. In the case under consideration, there are no periodic solutions in the vicinity of the resonance level. The most interesting case is when Eq. (11) has equilibrium states, i.e., when the condition

$$|B(I_{pq})| < \max_v|A_*(v, I_{pq})| \tag{12}$$

is satisfied. The resonance level $I = I_{pq}$ is then said to be partly passable.

Under condition (10), Eq. (6) may have limit cycles. In order to find them, one must construct the Poincaré-Pontryagin generating function.

Figure 1(a) shows the phase portrait of Eq. (6) under conditions (10) and (12), and $p = 3$. On the period $2\pi/3$, there is a single limit cycle (note that on the period 2π (which is the period of the unperturbed solution) there are three limit cycles). If the cycle lies outside the neighborhood of the separatrix loop of Eq. (11), then there is a corresponding two-dimensional invariant torus in the original system. Since the period of the limit cycle of Eq. (11) is of order $O(1/\mu)$, we then have a long-periodic beating regime in the original system (6) (the generatrices of the torus are of different order).

However, if the limit cycle lies in the neighborhood of the separatrix loop, then the two-dimensional invariant torus in the original system (1) is destroyed. The bifurcation scene in which the cycle is caught into the separatrix loop is shown in **Figure 1(b)**. Taking into account the nonautonomous terms, which were discarded in deriving Eq. (6), leads to the homoclinic structure. Such a structure is shown in **Figure 1(c)** for the Poincaré map with $p = 3$. Because of

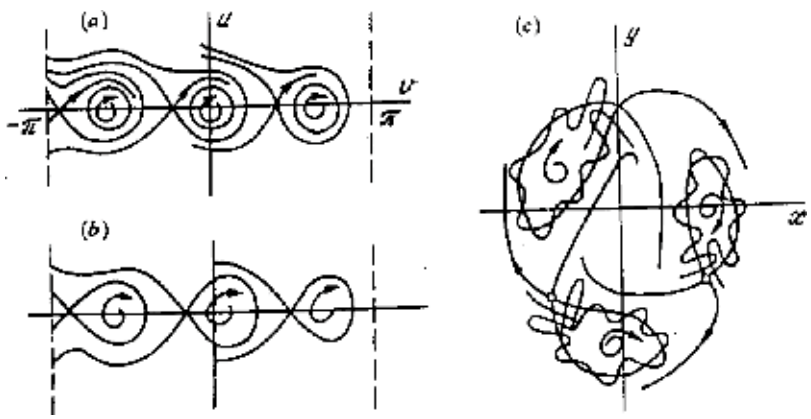

Figure 1. (a) Phase portrait of Eq. (6), (b) bifurcational case, and (c) Poincaré map for the initial system in case (b).

the presence of non-compact separatrices, in this case we merely have an irregular transition process.

Case II. Now, Eq. (6) always possesses equilibrium states, and we have the third kind of resonance zone, namely, an impassable zone. In order to better understand the structure of such a zone, we introduce in Eq. (6) the detuning γ between the level $I = I_{pq}$ and the level $I = I_0$, near which the autonomous system (2) has a limit cycle:

$$B(I_{pq}) = (dB(I_0)/dI)(I_{pq} - I_0) + O\Big((I_{pq} - I_0)^2\Big) \simeq \gamma\mu \tag{13}$$

Then, Eq. (6) can be rewritten as

$$\begin{aligned} du/d\tau &= A_*(v; I_{pq}) + \mu(\sigma(v; I_{pq})u + \gamma), \\ dv/d\tau &= bu. \end{aligned} \tag{14}$$

In Eq. (14) we change the variables from (u, v) to the action J and the angle L (in both the oscillatory and the rotational zones) and average the resulting system over the "fast" angular variable L. As a result, we arrive at the equation

$$\frac{dJ}{d\tau} = \mu b\mathbf{\Phi}(J)/2\pi,$$

where $\Phi(J)$ is the Poincaré-Pontryagin generating function [2] and it is discontinuous at $J = J_c$ when $\gamma \neq 0$. Here, J_c corresponds to the contour in the "unperturbed" system

$$\begin{aligned} du/d\tau &= A_*(v; I_{pq}), \\ dv/d\tau &= bu. \end{aligned} \tag{15}$$

formed by the saddle and two separatrix loops embracing the phase cylinder.

We shall therefore use Melnikov's formula [4] to determine the relative position of the separatrices which in the shortened system (15) constitute the contour formed by the outer separatrix loops:

$$\begin{aligned} \Delta &= \mu\Delta_1^{\mp} + O(\mu^2) \\ \Delta_1^{\mp} &= b\int_{-\infty}^{\infty} \big(\sigma_*(v_0; I_{pq}) + B_1(I_{pq})\big)u_0^2 d\tau \mp 2\pi\gamma. \end{aligned}$$

Here, v_0, u_0 is the solution of Eq. (15) on the contour consisting of the saddle and the outer separatrix loops. Setting $d = \max_v|\sigma_*(v; I_{pq})| = \|\sigma_*\|$, $a = |B_1(I_{pq})|/d$ we find from the formula for $\Delta_1^{\pm}$ that $\Delta_1^{\pm} = d(\alpha + \beta a) \pm 2\pi\gamma$, where

$$\alpha = b\int_0^{\infty} \overline{\sigma}(v_0; Ipq)u_0^2 d\tau, \beta = b\int_0^{\infty} u_0^2 d\tau, \overline{\sigma} = \frac{\sigma_*}{\|\sigma_*\|}.$$

From the condition $\Delta_1^{\pm} = 0$, we get

$$\gamma = \gamma^{\pm} = \mp d\big(\alpha + \beta a\big)/2\pi. \tag{16}$$

In system (14) the upper contour exists when $\gamma = \gamma^{+}$, and the lower contour when $\gamma = \gamma^{-}$. Eq. (16) defines two straight lines in the (a, γ) plane. They intersect each other at $(a^*, 0)$, where $a^* = -\alpha/\beta$. When $|a| > 1$ the function $\sigma\big(v; I_{pq}\big)$ is sign-preserving, and when $|a| < 1$ it is sign-alternating.

In virtue of Eq. (10), the second case is the most interesting. The case $|a| < 1$ is somewhat special since system (14) may then have limit cycles in both the oscillatory and rotational domains, which have no generating counterparts in system (2). Limit cycles in Eq. (14) can result from the following phenomena [5]: (a) from a degenerate focus, (b) from a separatrix loop (contour), and (c) from a condensation of trajectories. However, if the number of limit cycles does not matter, it suffices to consider the case when there is no more than one limit cycle in the oscillatory domain. Then, we can make a general conclusion on the change of qualitative dynamics of Eq. (14) under variation of the detuning. However, beforehand, we should study the problem for the case when f and g are trigonometric polynomials of degree N in φ. Then, A_* and σ_* are also trigonometric polynomials of degree $N_1 \le N$:

$$\begin{aligned} -bA_*\big(v; I_{pq}\big) &= \sum_{i=1}^{N_1} (a_i \cos(ipv) + b_i \sin(ipv)) \\ \sigma_*\big(v; I_{pq}\big) &= \sum_{i=1}^{N_1} (d_i \cos(ipv) + c_i \sin(ipv)). \end{aligned} \tag{17}$$

From the definition of functions $A_*(v)$ and $\sigma(v)$ (see Eq. (7)), it follows that, in general, different harmonics in the perturbation contribute to A_* and σ. This means that different harmonics can dominate in Eq. (17). We count only these main harmonics in Eq. (17) (for $A_* \Rightarrow 1$ and $\sigma_* \Rightarrow n$). We then derive from Eq. (6) the equation

$$z'' + \sin(z) = \mu\Big[(\cos(nz) + a)z' + \gamma\Big], \tag{18}$$

where $z = pv + \psi, \quad \psi = \arctan(b_1/a_1)$.

The generating function $\Phi(J)$ for Eq. (18) can be presented as [3]

$$\begin{aligned} &\Phi(J(\rho)) = \Phi^{(s)}(\rho) = aF_n^{(s)}(\rho) + F_0^{(s)}(\rho) \pm \delta_{2s} 2\pi\gamma \\ &F_0^{(1)}(\rho) = 16[(\rho - 1)\mathbf{K} + \mathbf{E}], F_1^{(1)}(\rho) = 16[(1 - \rho)\mathbf{K} + (2\rho - 1)\mathbf{E}]/3, \\ &F_0^{(2)}(\rho) = 8\mathbf{E}/\sqrt{\rho}, F_1^{(2)}(\rho) = 8[2(\rho - 1)\mathbf{K} + (2 - \rho)\mathbf{E}]/3\rho^{3/2} \end{aligned} \tag{19}$$

where $s = 1$ corresponds to the oscillatory domain and $s = 2$ to the rotational domain. $\mathbf{K}$, $\mathbf{E}$ are the complete elliptic integrals with modulus k ($\rho = k^2$). Note that $\rho = (1 + \tilde{h})/2$ in the oscillatory domain and $\rho = 2/(1 + \tilde{h})$ in the rotational domain, and $\tilde{h} = \tilde{h}(J(\rho))$ is the value of the energy integral of the equation $z'' + \sin(z) = 0$. Function $F_j^{(s)}(\rho)$ is the generating function defined by the

perturbation term $z^{'}\cos(jz)$. The plus in Eq. (19) corresponds to the upper half of the cylinder, the minus to the lower half, and δ is the Kronecker delta. This enables us to find all the bifurcation sets (except the one corresponding to a contractable separatrix loop) explicitly [6].

We shall first consider the case when $\gamma = 0$. In this case Eq. (18) is identical to the standard equation [2], and $\Phi(\rho)$ is continuous at $\rho = 1$. Thus, it determines the limit cycles up to the separatrix. This case was considered in **Figure 2(a–e)** that the rough topological structures are shown for $n = 1$. Note that the limit cycles can "disappear at infinity" only when $B_1 = 0$. This is impossible when Condition B is satisfied. **Figure 2(e)** shows the bifurcation when the limit cycle "clings" to the separatrix contour ($\Phi(\rho)$ has the simple root $\rho = 1$). **Figure 2(f)** shows the corresponding behavior of the invariant curves (separatrices) of the Poincaré map for the original system with $p = 3$. The neighborhood of the homoclinic contour is attracting. Moreover, a complicated structure exists in the neighborhood [7], and, consequently, we have a quasi-attractor, i.e., a nontrivial hyperbolic set, and stable points can exist in it.

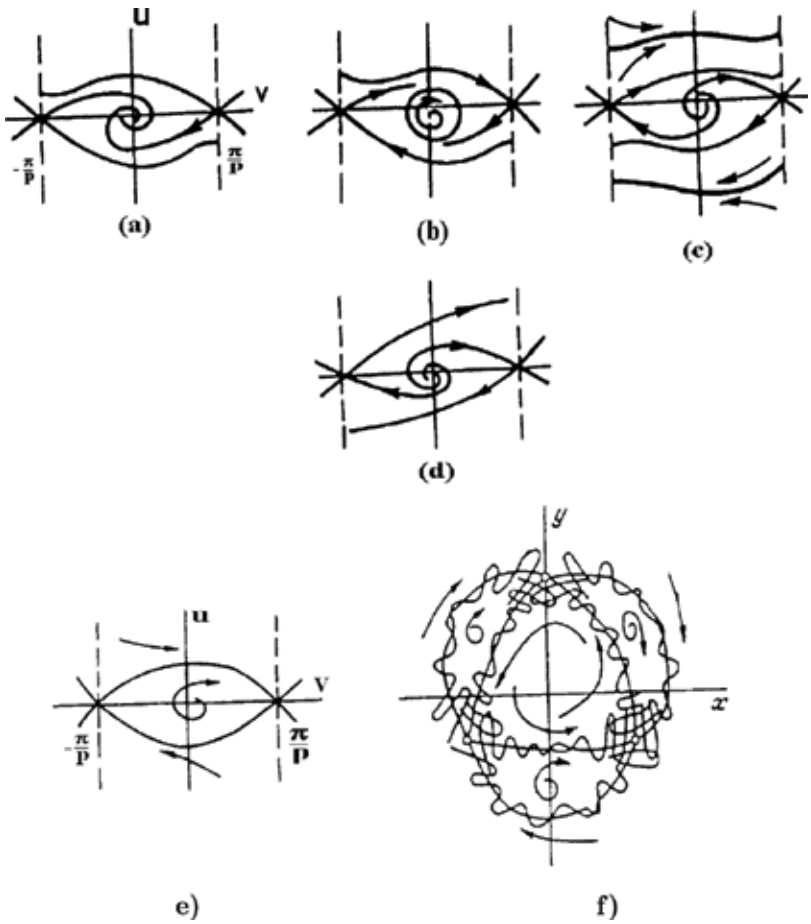

Figure 2. Phase portraits of Eq. (18) (a–e) and the Poincare map (f) for the case (e) and $p = 3$.

When $\gamma \neq 0$ the generating function $\Phi(\rho)$ is discontinuous at $\rho = 1$. The bifurcation of the cycle clinging to the separatrix must, therefore, be considered separately.

Using Melnikov's formula, we compute $\Delta_1^{\pm}$, which measures the split of the unperturbed separatrix for Eq. (18). One can see that equation $\Delta_1^{\pm} = 0$ is equivalent to $\Phi^{(2)}(1) = 0$. Then, using Eq. (19) and assuming (for concreteness) $n = 1$, we find the bifurcational values $\gamma^{\pm} = \mp 4(a + 1/3)/\pi$. When $\gamma = \gamma^{+} + O(\mu)$, we have a non-contractable separatrix loop lying in the domain $z' \geq 0$, and when $\gamma = \gamma^{-} + O(\mu)$, we have a loop in the domain $z' \leq 0$. From Eq. (19) we obtain the asymptotic formula, $\Phi^{(2)}(\rho) \simeq \pi(8a/\sqrt{\rho} + \sqrt{\rho} \pm 4\gamma)/2$, as $\rho \to 0$. This implies that the straight line $a = 0$ in the plane (a, γ) is singular. Furthermore, from Eq. (19) we find in the parametric form the line of the double cycles:

$$a = a_0(\rho) = -\left(F^{(2)}\right)'/\left(F^{(2)}\right)', \quad \gamma = \gamma_0(\rho) = \mp\left(F^{(2)}\left(F_n^{(2)}\right)' - \left(F^{(2)}\right)'F^{(2)}\right)/2\pi \cdot \left(F^{(2)}\right)',$$
$$\rho \in [0, 1], \quad \text{or} \quad \gamma = \gamma_0^{\pm}(a).$$

It is observed that the transformation of the phase portrait of Eq. (18) for $\rho \cong 1$ involves the creation of a contractable separatrix loop. By Condition *B*, we have $a \neq 0$, which implies that the saddle number is nonzero. The separatrix loop can, therefore, give rise to one limit cycle only [5]. The corresponding bifurcational set $\gamma_1^{\pm}(a)$ in the parameter plane can be found numerically.

We thus obtain a partition of the parameter plane (a, γ) into domains corresponding to different topological structures for Eq. (18), as well as the structures themselves (they are shown in **Figure 3**) for $n = 1$. The structures corresponding to cases 8–12 are not shown in **Figure 3**, since they can be obtained from structures 5, 6, 3, 2, and 14, respectively, by the directions of the coordinate axes.

Note that, along with a non-contractable separatrix loop, Eq. (18) has either a stable limit cycle, or a stable equilibrium state, or the stable "point at infinity." This means that no quasi-attractor can exist in the original nonautonomous system when $\gamma \neq 0$. Remark that the homoclinic structure exists for a small range of γ values $\left(|\gamma - \gamma^{\pm}| \simeq \exp(-1/\mu)\right)$.

Those limit cycles of Eq. (18) which do not lie in the neighborhood of the unperturbed separatrix contour correspond to the two-dimensional invariant tori in the original system (like in the case $B \neq 0$). Unlike when $B \neq 0$, two kinds of such tori may exist in Eq. (18) corresponding to the limit cycles in the oscillatory and rotational domains. The tori corresponding to the cycles in the rotational domain (with one exception) have no generating "Kolmogorov torus" in the perturbed Hamiltonian system, while the (asymptotically stable) tori corresponding to the limit cycles in the oscillatory domain are images of the tori occupying the next level in the hierarchy of resonances.

Remark The cases of odd and even n should be considered separately. When n is even, an unstable cycle clings to the separatrix loop. For odd n the same thing happens to a stable cycle. Only the case of odd n is therefore interesting when one studies the problem of existence of a quasi-attractor.

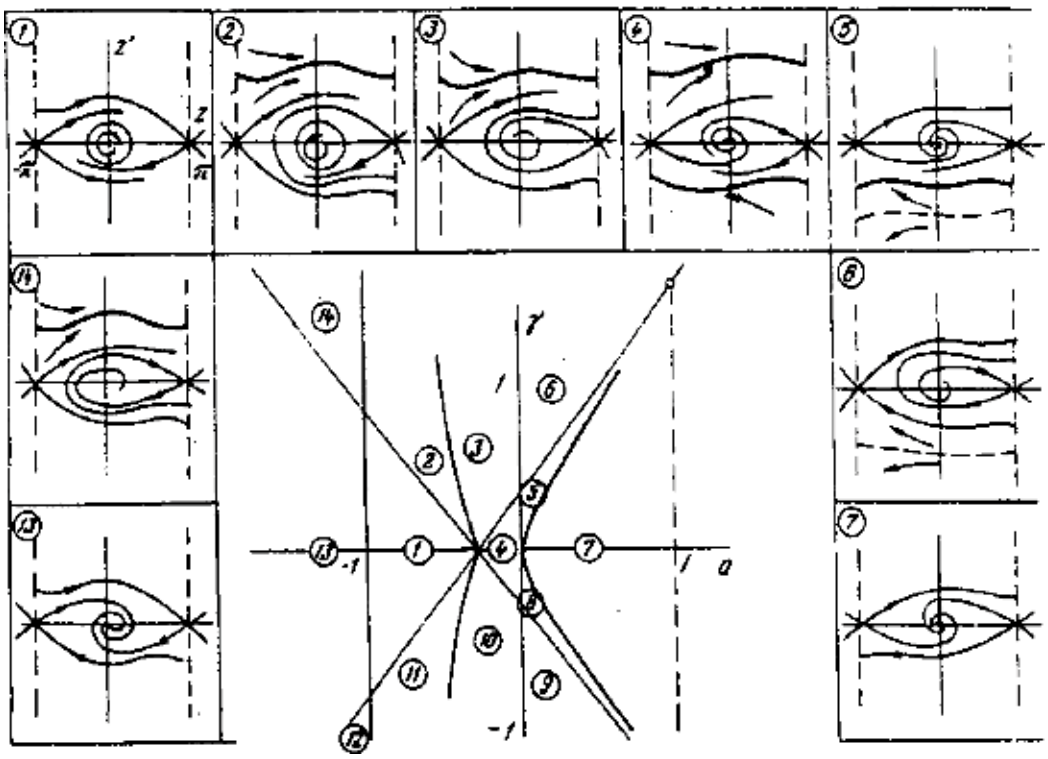

Figure 3. Bifurcation diagram and the corresponding rough phase portraits of Eq. (18).

According to the bifurcation diagram (**Figure 3**), it is convenient to break the case $|a| < 1$ into three sub-cases: (a) $-1 < a < a_*$, (b) $a_* < a < 0$, and (c) $0 < a < 1$, $a_* = 1/(1-4n^2)$. Let n be odd. Then considering the solutions on the original cylinder $\{v(\mathrm{mod}2\pi), u\}$, we derive the following theorem.

Theorem 2

There are μ_*, $\gamma^{\pm}(a)$, $\gamma_0^{\pm}(a)$, $\gamma_1^{\pm}(a)$, and a_* such that, if $|\mu| < \mu_*$ and n are odd, the following three intervals of a (in Eq. (18)) can be chosen: $1^\circ.\quad a \in (-1, a_*)$; $2^\circ.\quad a \in (a_*, 0)$; *and* $3^\circ.\quad a \in (0,1)$.

1. *Let $a \in (-1, a_*)$. Then, (1) when $\gamma > \gamma_1^+ > 0$, Eq. (14) has exactly one stable limit cycle (LC) in the rotational domain and no more than $p(n-1)$ LCs in the oscillatory domain (OD); (2) when $\gamma_1^+ < \gamma < \gamma^+$, there are p additional LCs in the OD, which are born from the separatrix loops at $\gamma = \gamma_1^+$; (3) when $\gamma = \gamma^+$, the stable LC in the rotational domain clings to the separatrix contour Γ_p^+ consisting of p saddles and their outer separatrices going from one saddle to another, while the "free" unstable separatrices approach an LC in the OD; (4) when $\gamma^- < \gamma < \gamma^+$, there are no LCs in the rotational domain and no more than pn LCs in the OD; (5) when $\gamma = \gamma^-$, there appears a separatrix contour Γ_p^- which consists of p saddles and their outer separatrices but has orientation and location different from those of Γ_p^+; (6) when $\gamma_1^- < \gamma < \gamma^-$, there are no more than pn LCs in*

the OD and one stable non-contractible LC; and (7) when $\gamma < \gamma_1^-$, Eq. (14) has one stable non-contractible LC which lies in the lower half-cylinder $u < 0$ and no more than $p(n-1)$ in the OD.

2. *Let $a \in (a_*, 0)$. Then, in the OD there are $p(n-1)$ LCs, and in the rotational domain, (1) when $\gamma > \gamma^-$, Eq. (14) has one stable LC for $u > 0$; (2) when $\gamma = \gamma^-$, a contour Γ_p^- appears; (3) when $\gamma^+ < \gamma < \gamma^-$, one stable LC exists on the upper half-cylinder $(u > 0)$ and one stable LC on the lower half-cylinder $(u < 0)$; (4) when $\gamma = \gamma^+$, a contour Γ_p^+ appears; and (5) when $\gamma < \gamma^+$, one stable LC exists for $u < 0$.*

3. *Let $a \in (0, 1)$. Then, there are at most $p(n-1)$ LCs in the OD, and in the rotational domain, (1) when $\gamma > \gamma^-$ and $u < 0$, Eq. (14) has one stable LC; (2) when $\gamma = \gamma^-$, a contour Γ_p^- appears; (3) when $\gamma_0^- < \gamma < \gamma^-$ and $u < 0$, there is a stable LC born from Γ_p^- and an unstable LC; (4) when $\gamma = \gamma_0^-$, the stable and unstable LCs merge together; (5) when $\gamma_0^+ < \gamma < \gamma_0^-$, no LCs exist; (6) when $\gamma = \gamma_0^+$, a semi-stable LC is formed for $u > 0$; (7) when $\gamma^+ < \gamma < \gamma_0^+$, one stable and one unstable LCs exist for $u > 0$; (8) when $\gamma = \gamma^+$, a contour Γ_p^+ is formed; and (9) when $\gamma < \gamma^+$, one unstable LC exists for $u > 0$.*

3. Example 1

Consider system (8) which is equivalent to the equation [3]

$$\ddot{x} + x + x^3 = \big(P_1 + P_2 x^2 + P_3 x \sin(\nu t)\big)\dot{x} + P_4 \sin(\nu t), \tag{20}$$

where P_i, $(i = 1, 2, 3, 4)$ are parameters. Here, we focus only on the effects which are due to the nonlinear parametric term $x\dot{x}\sin(\nu t)$. Let us assume $\nu = 4$. Then, for small $P_i (i = 1, 2, 3, 4)$ system (20) can have only two "splittable" resonance levels: $H(x, y) = h_{11}, H(x, y) = h_{31}$ and $h_{31} < h_{11}$. The corresponding autonomous system ($P_3 = P_4 = 0$) has at most one LC. The passage of this LC through the resonances under a change of parameter P_2 was considered in [2]. If this LC lies outside the neighborhoods of resonance levels $H(x, y) = h_{11}, H(x, y) = h_{31}$, then in the original nonautonomous system (20), there is a two-dimensional invariant torus T^2 corresponding to the cycle. There is a generating "Kolmogorov torus" in the Hamiltonian system ($P_1 = P_2 = P_3 = 0$).

A computer program was developed by the author for a simulation of Eq. (20). The results of such simulation are presented in **Figures 4–6**. In the numerical integration, the Runge-Kutta-type formulae are used with an error of order $O(h^6)$ per integration step h. In **Figure 4(a)** we present the Poincaré map for $P_1 = 0.0472$, $P_2 = -0.008$, and $P_3 = 0.018$, which determines the structure of the main resonance zone $(p = 1, q = 1)$. Along with the separatrices of the saddle fixed point S, a closed invariant curve encircling the unstable fixed point O is shown, which corresponds to a stable LC in the oscillatory domain of Eq. (6). This closed invariant curve appears for $P_3 \approx 0.014$ when the fixed point O loses its stability. As P_3 increases, so does the size of the closed invariant curve, and for $P_3 \approx 0.0487$ the curve clings to the separatrix of

http://dx.doi.org/10.5772/intechopen.79513

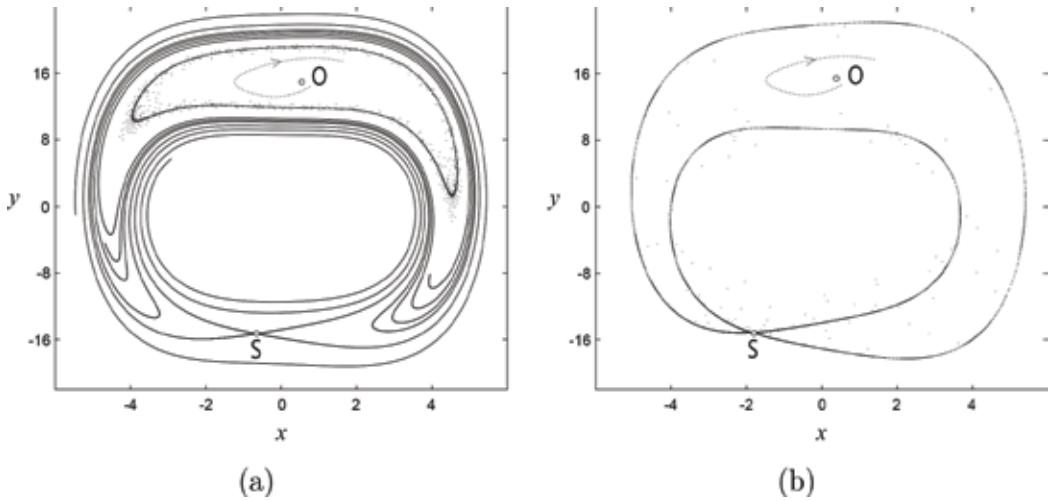

Figure 4. Poincaré map for Eq. (20) with $P_1 = 0.0472$, $P_2 = -0.008$, $P_4 = 2$, and $\nu = 4$ and (a) $P_3 = 0.018$ and (b) $P_3 = 0.0489755$.

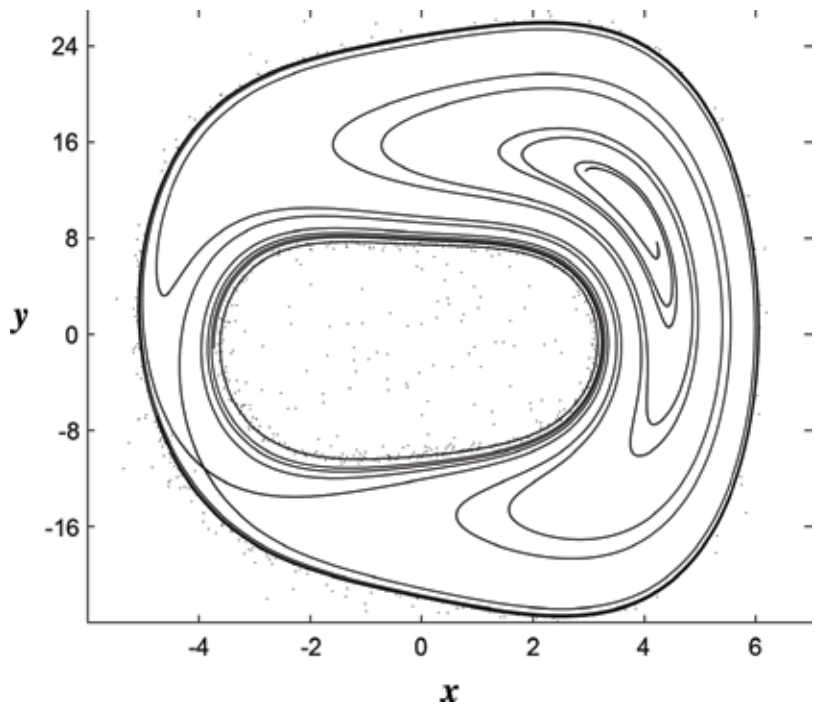

Figure 5. Poincaré map for Eq. (20) with $P_1 = 0.0472$, $P_2 = -0.008$, $P_3 = 0.15$, $P_4 = 2$, and $\nu = 4$.

the saddle point S, forming a contour (see **Figure 4(b)**). As P_3 increases further, two closed invariant curves appear, shown in **Figure 5** for $P_3 = 0.15$. The structural changes of the resonance zone observed in the experiment are in good agreement with the theoretical results for $\gamma = 0$. The observations for $\gamma \neq 0$ are consistent with the theory, too.

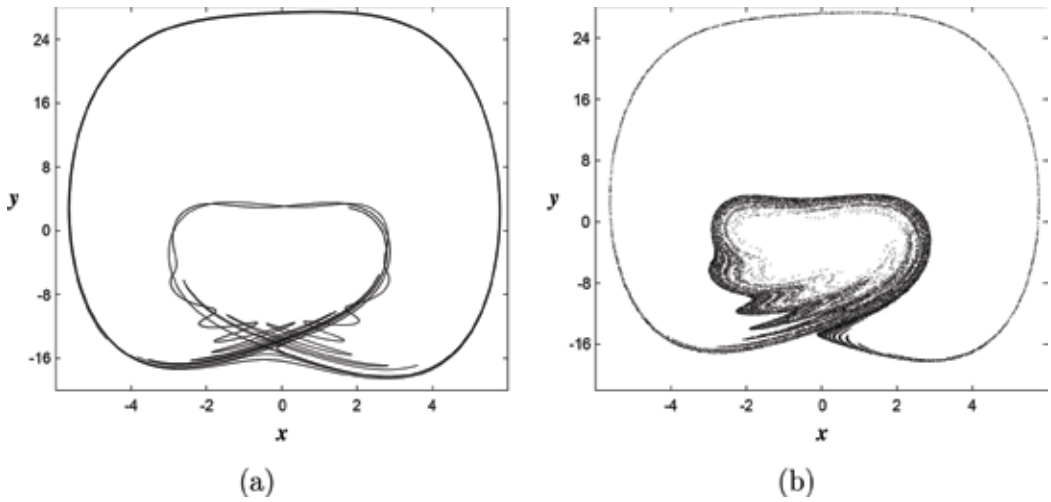

Figure 6. Poincaré map for Eq. (20) with $P_1 = 0.0472$, $P_2 = -0.008$, $P_3 = 0.0487$, $P_4 = 8$, and $\nu = 4$ (a) and quasi-attractor (b).

In the case presented in **Figure 6**, the transversal intersection of the separatrices of S cannot be detected visually. We, therefore, increased P_4 to obtain a better picture of the homoclinic structure. When $P_4 = 8$, the structure can be seen clearly (**Figure 6(a)**). The corresponding quasi-attractor is the only attracting set (**Figure 6(b)**). Stable periodic points with long periods can exist inside the quasi-attractor itself. However, they are extremely difficult to detect numerically.

4. Example 2

As opposed to Example 1, this one pursues a different goal, namely, to study the transition from the classical parametric resonance to the nonlinear resonance. One of the problems for which this can be done is that of the pendulum with a vibrating suspension.

The pendulum with vibrating suspension is a classical example of a problem in which a parametric resonance can be observed. A large number of publications (see, e.g., [8, 9]) are devoted to this problem. Other problems of this sort include the bending oscillations of straight rod under a periodic longitudinal force [10], the motion of a charged particle (electron) in the field of two running waves [11], etc. The parametric resonance in this kind of systems appears when a fixed point of the corresponding Poincaré map loses its stability and is, therefore, usually described by the linearization near this point.

It is interesting to study the behavior of a parametric system when the ring-like resonance zone is contracted into a point, i.e., to describe the bifurcations which occur in the course of transition from the plain nonlinear resonance to the parametric one. This paragraph is devoted to the solution of this problem in the case of a nonconservative pendulum with a vertically oscillating suspension.

http://dx.doi.org/10.5772/intechopen.79513

The motion of the pendulum with vertically oscillating suspension (under some simplifying assumptions) is described by the equation [13]

$$\ddot{x} + \sin x + p_1 \cos\beta t \sin x + p_2 \dot{x} = 0, \tag{21}$$

where p_1, p_2, β are parameters.

Let us now complicate the model even more and consider the equation

$$\ddot{x} + \sin x + p_1 \cos\beta t \sin x + \left(p_2 + p_3 \cos x\right)\dot{x} = 0, \tag{22}$$

with the phase space $\mathbf{R}^1 \times \mathbf{S}^1 \times \mathbf{S}^1$. The term $p_3 \dot{x} \cos x$ appears, for example, in the case of the pendulum in which the force of resistance is created by a vertical plate perpendicular to the plane of oscillations. Consider Eq. (22) when it is close to integrable, i.e., for small values of parameters $p_i (i = 1, 2, 3)$. Denote $p_i = \varepsilon C_i$, where ε is a small parameter. Then, the original Eq. (22) takes the form

$$\ddot{x} + \sin x = \varepsilon \left[C_1 \cos\beta t \sin x + (C_2 + C_3 \cos x)\dot{x}\right], \tag{23}$$

Eq. (23) in the conservative case, when $C_2 = C_3 = 0$, is considered in many publications. For instance, for small angles of the deviation x, the case $\beta \cong 2$ is studied in [8]. The criterion of resonance overlap is applied in [11] to estimating the width of the "ergodic layer." The existence of homoclinic solutions is discussed in [12] without the assumption on smallness of parameter ε.

Phase curves of the unperturbed mathematical pendulum equation are determined by the integral $H(x, \dot{x}) \equiv \dot{x}^2 - \cos x = h$, where $h \in (-1, 1)$ in the oscillatory domain and $h > 1$ in the rotational domain. The peculiarity lies in the way period τ depends on h in the oscillatory domain.

We have

$$\begin{aligned} &\tau(h) = 2\pi/\omega = 4\mathbf{K}(k), k^2 = (1 + h)/2, \ -1 < h < 1, \\ &\tau(h) = 2k\mathbf{K}, k^2 = 1/(1 + h), h > 1. \end{aligned} \tag{24}$$

Here, $\mathbf{K} = \mathbf{K}(k)$ is the complete elliptic integral of the first kind, k being its modulus. From Eq. (24) it follows that the period τ changes noticeably only for h close to 1, i.e., in the neighborhood of the separatrix. Therefore, small intervals of period τ, which determines the width of resonance zones, correspond to fairly large intervals of variable x.

4.1. Structure of resonant zones

In the investigation of the perturbed equation, we first focus on the structure of resonance zones in domains $G^1 = \{(x, \dot{x}) : -1 < h_- \le H(x, y) \le h_+ < 1\}$ and $G^2 = \{(x, \dot{x}) : H(x, y) \ge h_* > 1\}$. The resonance condition $\tau(h_{pq}) = (p/q)(2\pi/\beta)$, where p, q are relatively prime integers, determines the resonance levels of energy $H(x, y) = h_{pq}$.

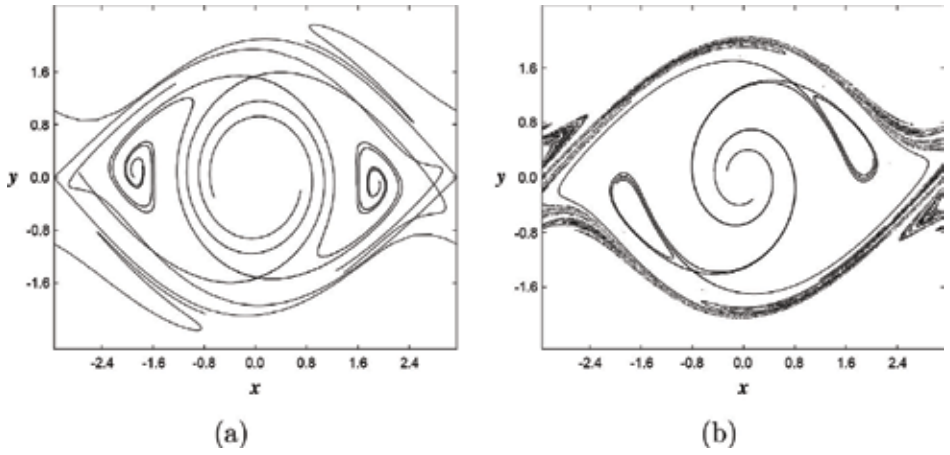

Figure 7. Invariant curves (separatrices) of Poincaré map for Eq. (22) with $p_1 = -0,1$, $p_3 = 0,1$, $\beta = 1.6$, and $p_2 = -0,07$ (a) with $p_2 \simeq -1/30$ (b).

The structure of individual resonance zones U_μ is described (up to the terms $O(\varepsilon^{3/2})$) by the pendulum-type Eq. (6). Since functions A_0 and σ have different forms in the oscillatory and rotational domains, we introduce the notations $A_0^{(s)}(v, h_{pq})$ and $\sigma^{(s)}(v, h_{pq})$, where $s = 1$ corresponds to the oscillatory domain and $s = 2$ to the rotational one.

In our case the divergence of the vector field of Eq. (23) contains no terms explicitly depending on t; hence, σ does not depend on v, i.e., $\sigma = \text{const}$.

The functions $A_0^{(s)}$ and $\sigma^{(s)}$ in an explicit form were obtained in [13]. It is also found that the width of the resonance zone decreases rapidly with the increase of p when $q = 1$.

A computer-generated picture of invariant curves of the Poincaré map for Eq. (22), with $\beta = 1.6$, is shown in **Figure 7**. In Eq. 21(a) a case of synchronization of oscillations in the subharmonic with $p = 2, q = 1$ ($p_1 = 0,1$, $p_2 = 0,07$, $p_3 = -0,1$) is shown, and in **Figure 7(b)**, a partly passable resonance with $p = 2, q = 1$ $(p_1 = 0,1, p_2 = 1/30, p_3 = -0,1)$ is shown. In the domain G^2 the synchronization of oscillations on the main resonance ($p = q = 1$) takes place.

4.2. Neighborhood of the origin

Denote $U_n = \{(x, y) : 0 \le H(x, y) \le C\varepsilon^{2/n}\}$ and substitute in Eq. (23):

$$x = \varepsilon^{1/n}\xi, \quad y = \dot{x} = \varepsilon^{1/n}\eta$$

As a result, we arrive at the system

$$\begin{aligned} &\dot{\xi} = \eta, \quad \dot{\eta} = -\xi + \varepsilon\left[C_1\xi\cos(\beta t) + (C_2 + C_3)\eta\right] + \varepsilon^{2/n}\xi^3/6 - \\ &-\varepsilon^{1+2/n}\left(C_1\xi^3\cos(\beta t)/6 + \xi^2\eta\right) + \ldots \end{aligned} \tag{25}$$

System (25) is defined in $D \times \mathbf{S}^1$ where D is a certain domain in $\mathbf{R}^2$. In the neighborhood $U_1\,(n=1)$, system (25) assumes the form

$$\dot{\xi} = \eta, \quad \dot{\eta} = -\xi + \varepsilon\left[C_1 \cdot \xi \cdot \cos(\beta t) + (C_2 + C_3)\eta\right] + O\left(\varepsilon^2\right). \tag{26}$$

By discarding in Eq. (26) the terms $O\left(\varepsilon^2\right)$, we arrive at the Mathieu equation with the extra term resulting from the viscous friction. It is clear that in the framework of a linear equation one cannot observe the (nonlinear) effects which accompany the transition from the nonlinear resonance to the parametric one. So, let us consider a wider neighborhood $U_2\,(n=2)$ of the origin. In Eq. (25) we discard the terms $O\left(\varepsilon^2\right)$ and, for the resulting system, consider the resonance cases when $\omega = 1 = q\beta/p$ (p and q being relatively prime integers). We then study the bifurcations pertaining to the transition from the parametric resonance to the ordinary one. We once again introduce the detuning $1 - q\beta/p = \gamma_1\varepsilon$. As a result, the system in question will be rewritten as

$$\begin{aligned} \dot{\xi} &= (q\beta/p)\eta + \gamma_1\varepsilon \\ \dot{\eta} &= -(q\beta/p)\xi + \varepsilon\left[C_1\xi\cos\beta t + (C_2 + C_3)\eta - \gamma_1\xi + \xi^3/6\right]. \end{aligned} \tag{27}$$

Now, we introduce the action (I) – angle (ϑ) variables. Since the unperturbed system is linear, the substitution has the simple form $\xi = \sqrt{2I}\sin\vartheta$ and $\eta = \sqrt{2I}\cos\vartheta$. In terms of this variables, system (27) will be written as

$$\dot{I} = \varepsilon F(I, \vartheta, \varphi), \quad \dot{\vartheta} = q\beta/p - \varepsilon R(I, \vartheta, \varphi), \quad \dot{\varphi} = \beta, \tag{28}$$

where $F = 2IG\cos\vartheta - \gamma_1\sqrt{2I}\sin\vartheta$, $R = G\sin\vartheta + \gamma_1\cos\vartheta/\sqrt{2I}$

$$G = C_1\sin\vartheta\cos\varphi + (C_2 + C_3)\cos\vartheta - \gamma_1\sin\vartheta + (I/3)\sin^3\vartheta.$$

Let us introduce in Eq. (28) the "resonance phase" $\psi = \vartheta - q\varphi/p$ and average the resulting system over the "fast" variable φ. As a result, we arrive at the two-dimensional autonomous system

$$\begin{aligned} \dot{u} &= \varepsilon[(C_1/2)u\sin 2v + (C_2 + C_3)u] \\ \dot{v} &= \varepsilon\left[(C_1/4)\cos 2v - u/8 - \gamma_1/2\right] \end{aligned} \tag{29}$$

when $p = 2$ and $q = 1$ and to the system

$$\begin{aligned} \dot{u} &= \varepsilon(C_2 + C_3) \\ \dot{v} &= \varepsilon\left(-u/8 - \gamma_1/2\right) \end{aligned} \tag{30}$$

when $p \neq 2$ and/or $q > 1$. As we know, $u = I + O(\varepsilon), \quad v = \psi + O\left(\varepsilon^2\right)$. From Eq. (29) and (30), it follows that (in our approximation) only one resonance with $p = 2, q = 1$ appears in the neighborhood U_2.

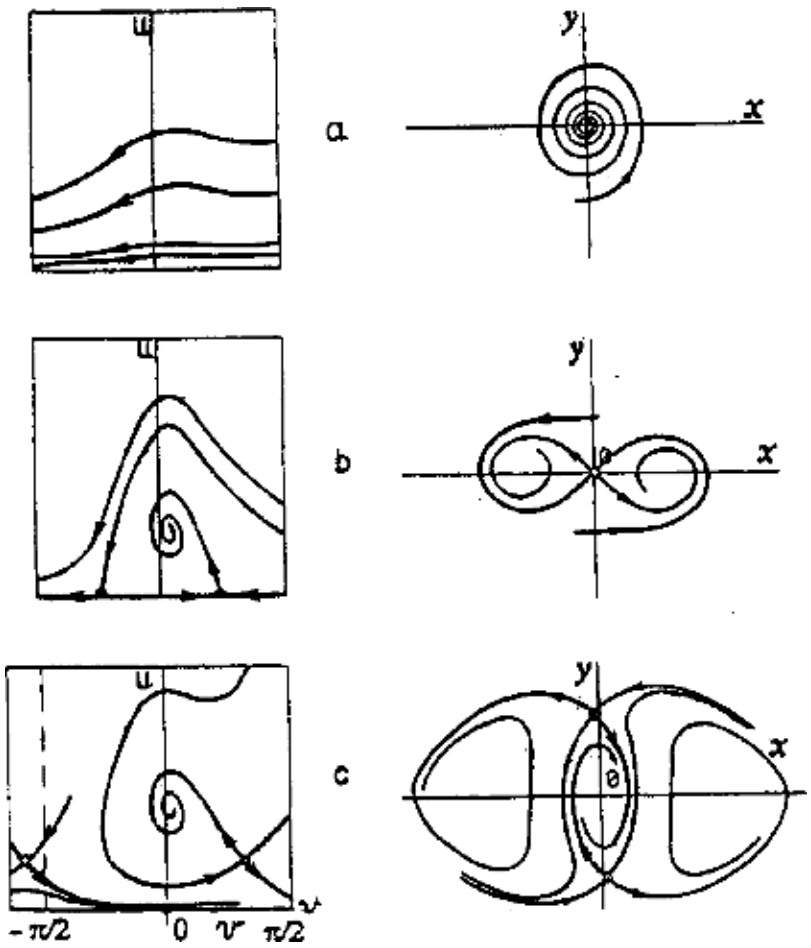

Figure 8. Phase portraits of system (29) with $C_2^2 + C_3^2 \neq 0$.

The investigation of system (29) when $C_2^2 + C_3^2 \neq 0$ for different values of detuning γ_1 presents no difficulty, because, according to the Bendixson criterion, there are no limit cycles. The most typical rough phase portraits are presented in **Figure 8** where, parallel with the phase portraits in the (u, v) plane, the corresponding phase portraits in Cartesian coordinates $(x, y = \dot{x})$ are shown. **Figure 8(a)** corresponds to the case when we have $\gamma_1 > \gamma_* > 0$, $\gamma_* = \sqrt{C_1^2 - 4(C_2 + C_3)^2}/2$, **Figure 8(b)** when $|\gamma_1| \leq \gamma_*$, and **Figure 8(c)** when $|\gamma_1| > \gamma_*$ and $\gamma_1 < 0$. In addition, in all three cases, we assume $C_2 + C_3 < 0$.

4.3. Conclusion

The number of splittable resonances is bounded, when $C_2^2 + C_3^2 \neq 0$. For the actual pendulum (Eq. (22)), when the small nonconservative forces are present, we, most likely, have one

resonance regime with $p = 2, q = 1$ in the oscillatory domain and the one with $p = 1, q = 1$ in the rotational domain.

In conclusion we make the following remarks on Eqs. (22) and (23).

1. The transition from **Figure 8(a)–(c)** corresponds to two period-doubling bifurcations, while the passage from the parametric resonance (**Figure 8(b)**) to the ordinary nonlinear resonance (**Figure 8(c)**) corresponds to the birth of two periodic (of period 2) saddle points and a node (focus) from a multiple saddle fixed point.
2. The bifurcation which involves the birth of a quasi-attractor (**Figure 7(b)**) in the neighborhood of the unperturbed separatrix is the most interesting one. It may take place at any magnitude of the external force (parameter C_1). It suffices to have $B^{(s)}(1) = 0, (C_2 = -C_3/3), \varepsilon(C_2 - C_3) < 0$, for example, $C_2 = -1/30, C_3 = 0.1, \varepsilon > 0$..
3. In the quasi-integrable nonconservative case, there appear no resonances with $q > 1$ and odd p in the oscillatory domain and no resonances with $q > 1$ and even p in the rotational domain.

Acknowledgements

This work was supported in part by the Russian Foundation for Basic Research under grant no. 18-01-00306, by the Russian Science Foundation under grant no. 14-41-00044.

Author details

Albert Morozov

Address all correspondence to: morozov@mm.unn.ru

Lobachevsky State University of Nizhny Novgorod, Russia

References

[1] Morozov AD, Shil'nikov LP. On nonconservative periodic systems similar to two-dimensional Hamiltonian ones. Pricl. Mat. i Mekh. (Russian). 1983;**47**(3):385-394

[2] Morozov AD. Quasi-conservative systems: Cycles, resonances and chaos. World Scientific Series on Nonlinear Science Series A. 1998;**30**. http://www.worldscientific.com/worldscibooks/10.1142/3238

[3] Morozov AD. Resonances and chaos in parametric systems. Journal of Applied Mathematics and Mechanics. 1994;**58**(3):413-423

[4] Melnikov VK. On stability of a center under periodic in time perturbations. Work of the Moscow Mathematical Society. 1963;**12**:3-52

[5] Andronov AA, Leontovich EA, Gordon II, Maiyer AG. The Theory of Bifurcations of Dynamical Systems in a Plane. Moscow, Russia: Publ. Nauka; 1967

[6] Morozov AD, Dragunov TN. Visualization and the Analysis of Dynamic Systems. Moscow-Izhevsk: Publ. Institute of Computer Researches; 2003

[7] Shil'nikov LP. About the Poincaré-Birkgof problem. Mathematical Reviews. 1997;**174**(3): 378-397 Russian

[8] Struble RA. Oscillations of a pendulum under parametric excitation. Quarterly of Applied Mathematics. 1963;**21**(2):121-131

[9] Struble RA, Marlin JA. Periodic motion of a simple pendulum with periodic disturbance. Quarterly Journal of Mechanics and Applied Mathematics. 1965;**18**(4):405-417

[10] Bolotin VV. Dynamic Stability of Elastic Systems. Moscow: Gostehizdat; 1956 (Russian)

[11] Zaslavsky GM, Chirikov BV. Stochastic unstability of nonlinear oscillations. Uspekhi Fizicheskikh Nauk (Russian). 1971;**105**(1):3-39

[12] Cherry TM. The asymptotical solutions of the analytical hamiltonian systems. Journal of Differential Equations. 1969;**4**(2):142-156

[13] Morozov AD. The problem of pendulum with an oscillating point of suspension. Journal of Applied Mathematics and Mechanics. 1995;**59**(4):563-570

Chapter 6

Mechanical Perturbations at the Working Electrode to Materials Synthesis by Electrodeposition

Baudel Lara Lara, Arturo Fernández Madrigal,
Lizbeth Morales Salas and
Alejandro Altamirano Gutiérrez

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.78544

Abstract

Applying mechanical perturbations at the working electrode during the electrodeposition process is a novel strategy for materials synthesis that has been used for Cu(In,Ga)Se_2 (CIGS) thin film synthesis. A mechanical perturbations strategy was applied during one-step electrodeposition, and the results are compared with the traditional one-step electrodeposition where no mechanical perturbations were applied. In both cases, a potentiostatic mode was employed, where DC potential is applied to the working electrode with respect to the reference electrode; the potential is regulated by the current at an auxiliary electrode. The CIGS films obtained from both strategies were analyzed as electrodeposited and after being annealed in a selenium atmosphere. The annealed film morphology obtained with the potentiostatic mode plus periodical mechanical perturbations was denser and more compact than the film without mechanical perturbations. Using contour lines, the morphology evolution and mass transport distribution on the working electrode during the electrodeposition process are explained.

Keywords: mechanical perturbations, electrodeposition, thin film, morphology

1. Introduction

Advance of new materials is an important issue in different areas of the technological development; however, the challenges are still diverse and important. A necessary aspect of human survival is the use of renewable energies, especially the solar energy. A technology that can convert the solar radiation in electricity directly is the use of solar cells. Different methods for

materials synthesis have been widely investigated for solar energy conversion. These methods can be classified into physical and chemical. Electrodeposition is a chemical method that has been used to obtain metallic or semiconducting films on substrates, with the aim of protecting the surface against the oxidation and corrosion, giving a better esthetic appearance, and providing some mechanical and electrical characteristics, different to the base material, to improve its physical properties. Electrodeposition has been considered for solar cell applications for a long time [1], with a considerable potential for the fabrication of a low-cost thin film for solar cells [2]. It is simple, versatile, and economical as compared to physical methods, such as high vacuum processes. Some of its advantages are requiring less capital investment, saving raw material, application on irregular surfaces, and industry scalability potential. It has been used for materials synthesis for perovskite [3], Cu_2ZnSnS_4 [4], and $CuIn_xGa_{(1-x)}Se_2$ (CIGS) solar cell [5]. It is also used for nanoparticle synthesis for solar collectors [6] and to develop technology in the energy storage [7].

The electrodeposition method is considered difficult; perhaps because the electrodeposition process is affected by many variables. Some of them are concentration of the solution, solution temperature, pH, working electrode potential, working electrode resistivity, and distance between electrodes. During the electrodeposition process, there are phenomena that affect the film growth, among them, the diffusion layer, the depletion region, and the natural flow by convection. For such reasons, the electrodeposition method is still under investigation to improve the material quality.

2. Conventional electrodeposition

A conventional three-electrode electrolytic cell connected to potentiostat equipment is used for material synthesis investigation in solar cell applications by electrodeposition. A direct current (DC) potential is applied to the working electrode (WE) with respect to the reference electrode (RE). The DC potential in the WE is regulated in a closed-loop control system by adjusting the current at an auxiliary electrode (AE). According to the electrolytic solution, different materials can be deposited on the WE. Synthesizing a material by the electrodeposition technique consists of finding the values of the variables that produce a film growth with the desired characteristics. Thin films of materials with homogeneous growth and compact morphology are desirable. If the material to be formed consists of two or more elements, there are diverse electrodeposition alternatives, principally (1) electrodeposition of elements in layers, followed by annealing, to get the desired structure, and (2) electrodeposition of elements in a simultaneous way, also followed by annealing, to increase crystallinity. The latter alternative is known as a one-step electrodeposition in which the electrochemical conditions must be met in such a way that the material composition is homogeneous from the first instant of formation. The advantage of the one-step electrodeposition alternative is that the film that is obtained is electro-crystallized. In addition, a single electrolytic solution with ions of the precursor elements is used, while in the electrodeposition by layers, an electrolytic solution is used for each layer.

The basic modes of crystal growth on a WE under electrochemical conditions have been established; the electrodeposition takes place at the electrode-electrolyte interface under the

http://dx.doi.org/10.5772/intechopen.78544

influence of an electric field [8]. A scheme of load distribution as well as a simple electrical model of a three-electrode electrolytic cell is represented in **Figure 1**. Charge distribution in the electrode-electrolyte interface is analogous to charge distribution in a capacitor which is called the electrical double layer. The electric field lines are defined only at the interface, where C_{WE} and C_{AE} are the equivalent capacitances between the WE and AE and the solution, respectively. R_{SOL} is the solution electric resistance. In a three-electrode-electrolytic cell, the electrode-solution interface is principally important in the WE and AE. Other detailed electric models [9] for the electrode-solution interface suggest that the interface is affected by the capacitance of the double layer, the resistance in the charge transfer zone, and the impedance due to adsorption and mass transport, and the parameter of the electrical model, mentioned earlier, can be obtained by electrochemical impedance [10].

In an electrodeposition process, the charge, which consists of electrons, is considered to be evenly distributed on the WE surface. When M^{2+} ions are present in the solution, the electrochemical of $M^{2+} + 2e \rightarrow M$ is carrier out on the WE. When more than one type of ions must be reduced, the transport mechanism of the ions until the load transfer zone plays an important role. It is assumed that the ratio at which the ions are consumed by the charge transfer reaction is equal to the ratio at which the ions arrive at the charge transfer zone.

At the beginning of the electrodeposition process, the first nuclei grow on the WE surface. At that time, the solution around the nuclei is depleted of ionic species, and the only factor that influences the ion movement is diffusion [11], through which the depleted zones around the

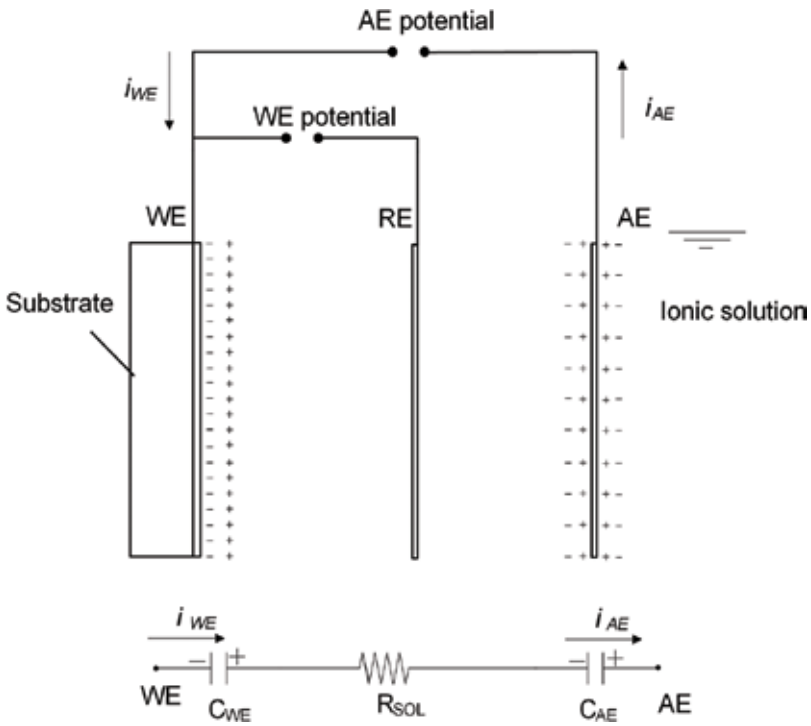

Figure 1. A diagram of a simple electrical model of a three-electrode electrolytic cell.

nuclei start to propagate throughout the WE surface; as the electrodeposition time elapses, these zones begin to overlap [12]. **Figure 2** shows a schematic representation of the ion movement distribution on the WE surface at the initial stage of the film formation. The arrow lines

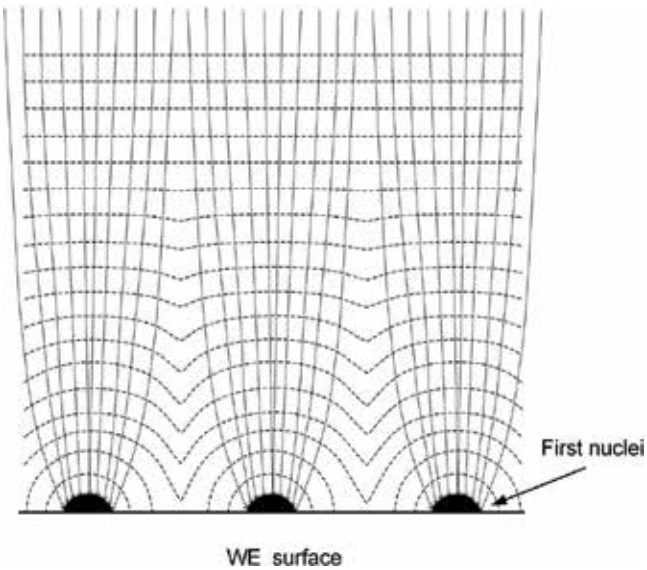

Figure 2. A schematic representation of the mass transport distribution at the initial stage of the film formation by electrodeposition (modified from [12]).

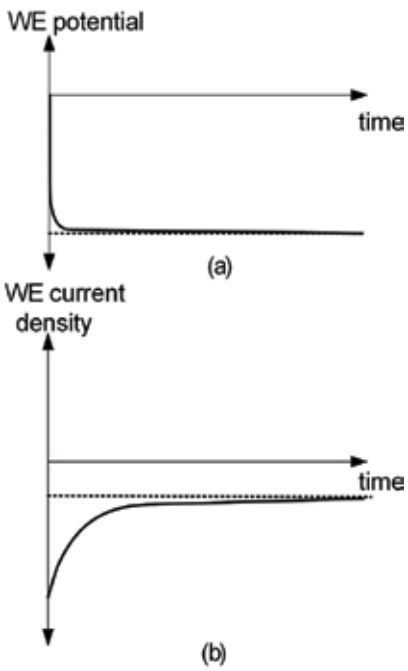

Figure 3. A graphic representation of electrodeposition signals versus time obtained during the electrodeposition process: (a) WE potential and (b) WE current density [13].

show the ion movement from the bulk solution to the WE vicinity. In this region, the ions concentration is different from their value in the bulk solution; this region is known as the diffusion layer.

When the DC potential is applied at the WE, the current density versus time shows a transitory stage evolution, followed by a stage where the current variation is lower than the transitory stage. Also, the potential has a transitory stage; after that, the WE potential reaches the set value. With acquisition data software, the voltage and current signals can be plotted as time function, as it is shown in the graph in **Figure 3**. According to the above, the electric model of the three-electrode-electrolytic cell represents an average model of the electrical phenomena that take place in the electrode-solution interface during the electrodeposition process.

3. CIGS thin film by electrodeposition

The semiconductor CIGS is a promising absorber in thin film solar cells, due to their direct band gap and large optical absorption coefficient. Small-area CIGS solar cells with efficiencies reaching 22.6% have been built with this semiconductor synthesized by the high vacuum deposition method [14]. A compact absorber morphology is a quality indicator to obtain high performance CIGS solar cells. To improve solar cell efficiency, the CIGS absorber should have, in cross view, large grains extending from the back to the front [15]. The CIGS absorber has the highest potential to develop large-scale solar cells. Employing high vacuum deposition method, high efficiencies have been achieved; however, economical deposition methods with the possibility of implementing in large area still need to be developed [16, 17]. It is considered that the CIGS absorber can be synthesized in a large area and with compact morphology employing the electrodeposition method. However, when CIGS solar cells have been built by synthesizing the absorber by electrodeposition, solar cell efficiencies of 11.3% for a one-step electrodeposition [18] and 14.17% for layers electrodeposition [19] have been achieved. The low efficiency is attributed to lack of absorber quality when it is obtained by the electrodeposition method, usually associated with the morphology. Microcracks have been identified in the CIGS film obtained by a one-step electrodeposition [20], which is one reason why relatively low efficiencies are obtained in the CIGS solar cell by electrodeposition.

To develop the CIGS solar cells on a large scale by electrodeposition, there are still aspects that must be investigated. The major challenges and required strategies have been identified. Among them, (1) the precise control of film stoichiometry (optimization of Ga content and Ga distribution), (2) novel deposition strategies, (3) understanding on the mechanism of Ga incorporation, and (4) establishing the strategy that allows electrodepositing the semiconductor with homogeneous composition and uniform morphology throughout the film [16, 21].

In several works, electrodeposition strategies for thin films synthesis have been used. It has been established that by using a pH-regulating solution, stability is provided to the electrodeposition process. No oxides or hydroxides are obtained in the solution, and it is possible to incorporate a higher percentage of gallium in the film [20]. In the first stages of Cu-In-Se on Mo-coated glass by electrochemical deposition, the first nuclei are made of a copper-rich Cu-Se without indium and the nucleation is developed by a quasi-instantaneous three-dimensional

nucleation [22]. In $CuInSe_2$ (CIS) one-step electrodeposition, it has been established that the Cu-Se phase is formed at a low potential, and a reaction path has been established as a function of the potential. The Cu-Se phase acts as a nucleation site for indium incorporation [23, 24]. The CIS film morphology deposited at various potentials has been analyzed [23]. At low polarizations between −0.4 and − 0.5 V, platelets characteristic of the Cu-Se were observed; when the polarization increased, the morphology was nodular. The mechanisms of Ga to CIS incorporation also have been established. It is incorporated as gallium selenide and GaO_3 [25]. The CIGS film morphology obtained by the one-step electrodeposition with potentiostat mode has been described as nodules with a cauliflower-like growth [22, 26]. The as-electrodeposited CIGS film morphology is strongly influenced by the bath composition. Microcracks in the films have been observed when the films were deposited at low concentrations of $CuCl_3$, $InCl_3$, and $GaCl_3$ salts and at high concentrations of H_2SeO_3 [27].

Many studies have examined ways of improving the CIGS film morphology by a one-step electrodeposition. The effect of sodium sulfamate as a complexing agent on the film morphology was evaluated [28]. An improvement on CIGS thin film morphology was obtained when a short electrode pretreatment of a 1-min deposition at −0.5 V was carried out prior to deposition of the film [29]. The pulse electrodeposition process can produce a CIGS film that is more smooth, compact, and homogeneous than the one deposited by the DC potential electrodeposition [30]. Electrochemical studies in CIGS electrodeposition, generally, use an electrochemical cell with electrodes suspended vertically. However, an electrochemical cell with electrodes in a horizontal position has advantages over a cell with vertical electrodes, principally because the ion transport mechanism as well as the natural flow by convection allows a better uniformity on the WE surface; in this way, the composition is homogeneous through the film [31].

3.1. Characterization of CIGS films obtained by electrodeposition

An electrochemical cell system of three horizontal electrodes was installed with a scheme like the one shown in **Figure 4**. A glass substrate covered by an Mo film (1 μm of thickness and $4x10^{-4}$ Ω cm of resistivity) was the WE. The RE and AE were made of a platinum mesh. The CIGS films were electrodeposited by applying −1.0-V DC potential to the WE versus the RE, employing an electrolytic solution with copper, indium, gallium, and selenium ions. At the start, it was the electrodeposition process, where a stage of nucleation and electrocrystallization of the CIGS film on the WE electrode was obtained. After the electrodeposition process, the WE with the CIGS film was removed from the electrochemical bath, rinsed with deionized water, and placed vertically for drying. Although copper, indium, gallium, and selenium ions have different reduction potential, a situation that complicates a simultaneous ED process, the CIGS films have been obtained with the composition ratios of Ga/(In+Ga) = 0.31 and Cu/(In+Ga) ≈ 0.9, close to those reported in the high efficiency cells [32, 33]. The film composition was measured by atomic emission spectroscopy (ICP-AES). The current evolution indicates that the steady-state value can be reached after 5 min, with a limiting current density of ≈ 1 mA / cm^2. The WE surface changes during the film formation affect the limiting current density in the steady state in such a way that it decreases with very slow dynamics. The above also indicates that diffusion layer thickness increases. During the steady stage, the reaction at the WE is affected by the

transport of the chemical species from the bulk solution to the charge transfer zone. By increasing the electrodeposition time, the film obtained is more rugged and darker in color; this is due to the lack of ions near to WE and to the increase of the diffusion layer thickness. The stirring of the solution is desirable since it enhances ion transport to the substrate and decreases the thickness of the diffusion layer [34]. However, the agitation method of stirring for a laboratory-scale deposition leads to gradients of thickness in the flow direction of the electrolyte [35].

The CIGS films characterized by scanning electron microscope (SEM) are shown in **Figure 5**. **Figure 5(a, b)** shows micrographs of surface and cross section. In both cases, the morphology consists of vertical nodules with a well-defined boundary between them. Some nodules are larger than others, which apparently have stopped growing. The stunted nodules increase the boundary between the nodules that have a greater growth. **Figure 5(c)** shows the surface morphology of the vertical nodules with a cauliflower-like growth. The surface morphology among the nodule boundaries is shown in **Figure 5(d)**. The film morphology in the nodule has differences with the one that exists in the boundary. Apparently, the film formed between boundaries is less compact than those formed in the nodule. In general, the CIGS films that were obtained through the one-step electrodeposition are not very compact and have a low crystalline structure, so that they do not have the properties to be used in solar cells. The principal morphology consisted in groups of atoms forming the cauliflower-like growth. The annealing process in a selenium atmosphere is necessary to transform the as-electrodeposited film into a more crystalline, with large grains and with compact morphology.

The CIGS thin films that were subjected to an annealing process in a selenium atmosphere are shown in **Figure 6**. The selenization temperature was 550°C for 180 min. **Figure 6(a,b)** shows the surface and cross-section micrographs. In the micrographs, there is evidence that the nodules are of different length. On increasing the deposition time, some nodules continue

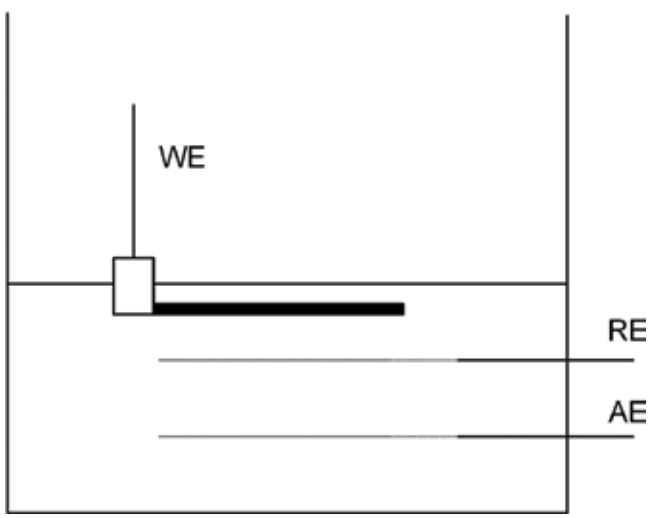

Figure 4. A diagram of an electrolytic cell with three horizontal electrodes.

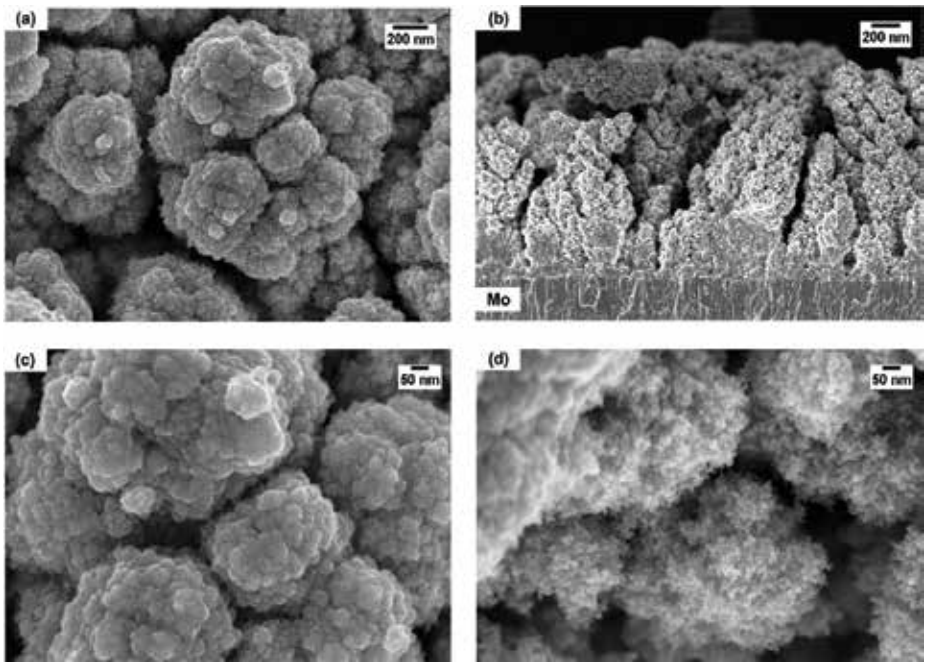

Figure 5. Micrographs of the CIGS film that has been electrodeposited in a conventional mode: (a) surface, (b) cross section, (c) nodule with a cauliflower-like growth, and (d) morphology in the nodule boundary [13].

to grow and others stop growing. The cross-section film micrograph shows the Mo layer and over it, a CIGS film with a compact morphology with 300 nm of thickness; this layer is also evident from the as-electrodeposited film shown in **Figure 5(b)**. It was noticeable that the compact layer is due to the initial growth when the current density is in a transitory state and the diffusion layer is thin. Over the compact CIGS film, there are only formations of isolated nodules of different sizes with very large boundaries between them; in this stage of formation, the current density and the mechanism of mass transport are not locally uniform. In order to reduce the activation energy and grow large grains during the annealing process, a Cu-rich film was prepared. The film composition ratios were Ga/(In+Ga) = 0.29 and Cu/(In+Ga) = 1.18. The Cu content of the film determines the activation energy for grain boundary motion. It has been determined that by increasing the Cu content of the film from 17.9 to 25.7%, the activation energy decreases from 3.5 to 3.0 eV [36]. The micrographs of annealed films are shown in **Figure 6(c, d)**. From these micrographs, it can be noted that there is also a compact CIGS film over the Mo film, which shows that in the copper-poor and copper-rich films, the films are compact in the first stage of growth, up to a thickness of 300 nm. A nonuniform grain growth is identified. There is only a grain growth in the boundaries indicating that the kinetic of grain growth during selenization process in the nodule boundaries was different, which is believed to be caused by the non-homogeneity in the film composition originated by the nonuniformity of the current density during the one-step electrodeposition process. Copper-rich films were formed in the nodule boundaries. In this way, the atomic composition in the films is not uniform, the nodules are copper-poor, and the boundary nodules are copper-rich. That is, the film locally will have different electrical, structural, and optical characteristics.

http://dx.doi.org/10.5772/intechopen.78544

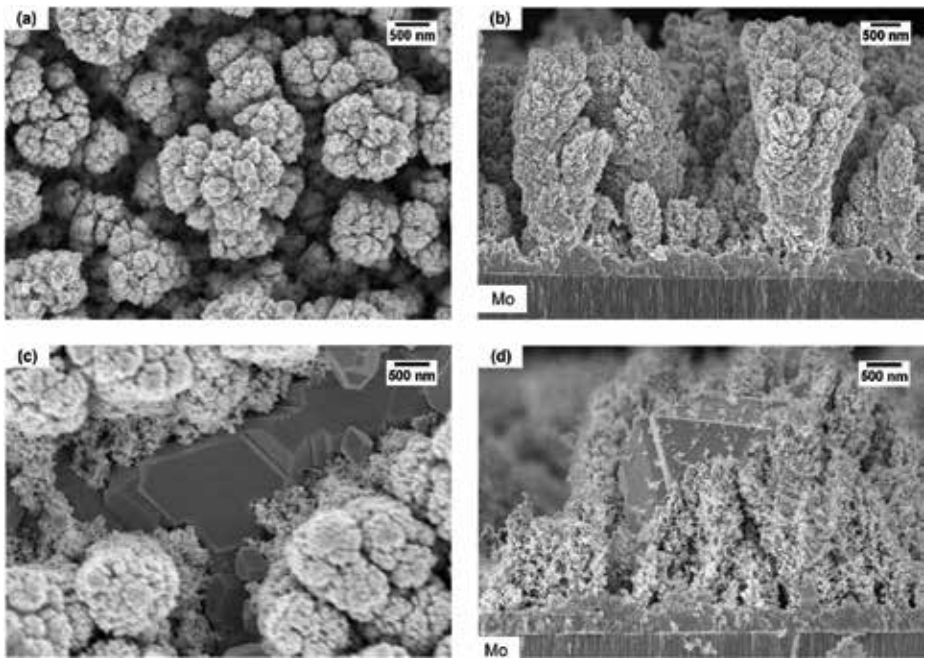

Figure 6. Micrographs of surface and cross section of the CIGS films with (a, b) composition ratio Cu/(In + Ga) = 0.9 and (c, d) composition ratio Cu/(In + Ga) = 1.18 [13].

4. Conventional electrodeposition plus periodical perturbations

In the electrodeposition theory, it is assumed that the WE surface is homogeneous so that the current density in the macroscopic level is uniformly distributed over the WE surface [8]. However, at the microscopic level, if the surface of the WE is considered as a surface with roughness, there will be a greater electric field strength in the peaks than in the surface valleys, as shown in **Figure 7(a)**, where the WE roughness has been amplified. Thus, the electrochemical kinetic is affected. For this reason, the electric load will be concentrated in the crests of the WE. With the formation of the first nodules, the WE roughness increases in such a way that the current density and therefore the mass transport mechanism are concentrated at the nodules. The foregoing has been observed in other studies, where it has been determined that the nonuniformity in the local current densities can exist even when the macroscopic current distribution over a given surface is completely uniform [34]. Assuming that there is a direct relationship between the load transfer ratio and the current density that is demanded during the electrodeposition process, the load transfer process can be analyzed using the current density. The points of greatest intensity of the electric field are the crests of the WE. In this way, they produce a greater current density during the ion reduction process, in such a way that, in the crests, the growth of the film originates grains that grow perpendicularly with respect to the WE. In the valleys, the current density is lower, and therefore, the density of ions is reduced and the growth speed of the film is slower. From the previous results, it can be established that as a consequence of nonuniformity in the local current densities through the WE, due to the diffusion layer growth, the CIGS morphology consists of isolated nodules

with a cauliflower-like growth. It is evident after the transitory stage in the electrodeposition current density, and this is more noticeable when the electrodeposition time increases, a representation of isolated nodules is shown in **Figure 7(b)**.

The contour lines are used in topographic maps to represent points of the same elevation. Here, we use it to represent the CIGS film morphology as shown in **Figure 8**. The contour lines were traced according to the film morphology shown in **Figures 5** and **6**. The contour lines represent the CIGS film morphology of the same thickness. According to the micrographs, the CIGS films morphology is not uniform through the WE surface, being a function of the current density distribution during the electrodeposition process. Thus, the contour lines also represent the mass transport mechanism during the electrodeposition process. **Figure 8** shows the contour lines that represent the electrodeposition process evolution. The highest mass transport density and the nodule formation are represented in zones with a dark color while less mass transport density and the boundary layer formation between nodules are represented with a light gray color. When the electrodeposition process begins, a large number of nodules grow randomly distributed. The nodules distributions, which can be appreciated after the current transitory stage, are represented in **Figure 8(a)**. When the first nodule has a cauliflower-like growth, the mass transport mechanism is concentrated in them; however, not all nodules grow at the same rate and eventually some stop growing. When the deposition time elapses, the nodules become more and more isolated and the boundary between them grows, because the mass transport mechanism is concentrated at very isolated points; this is represented as contour lines in **Figure 8(b, c)**. The above causes the CIGS film morphology to appear as isolated nodules with a high roughness, and this is evident when the electrodeposition time increases. In previous studies, it has been shown that in the CIGS film synthesized by electrodeposition, the composition is related to the current density; if

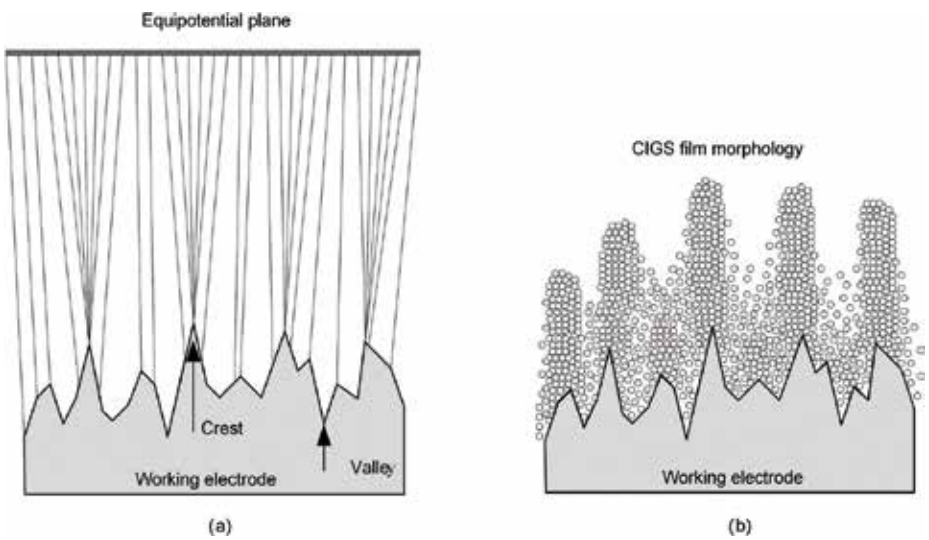

Figure 7. Growth model of CIGS films, obtained by the one-step electrodeposition: (a) effect of the WE roughness on the distribution of the electric field and (b) effect of the distribution of the electric field on the film morphology.

http://dx.doi.org/10.5772/intechopen.78544

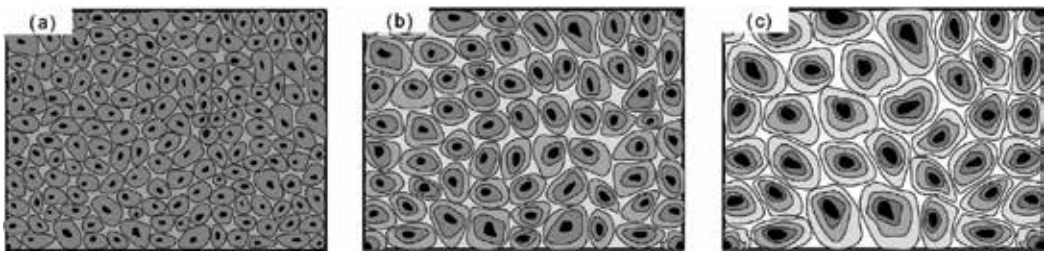

Figure 8. Contour lines that represent the CIGS film morphology, mass transport, and current density distribution in three different stages of the electrodeposition process [13].

the current density is high, the CIGS film composition is copper-poor. On the other hand, if the current density is low, the CIGS film composition is copper-rich [31]. Therefore, the CIGS film has a composition that varies according to the contour line; the zones with a dark color are cooper-poor while the zones with a light gray color are copper-rich. The above is evident in the surface morphology of CIGS films that were subjected to an annealing process as shown in **Figure 6(c, d)**. At the initial stage of the electrodeposition process, the ratio of ion reduction in the WE surface is much greater than the speed at which the ionic species arrive at the load transfer zone; for this reason, the diffusion layer increases up to maintain an electric load balance. In this way, the zone of the depletion varies according to the contour plot. At this microscopic level, the diffusion layer is not uniform along the working electrode. It acquired this principal shape after the transient of the electrodeposition current density. At the macroscopic level, experimental results [37] revealed that the electrochemical kinetic behavior of CIGS thin films is strongly influenced by the electrical double layer existing between the substrate. With an increase in the electrodeposition time, the kinetic behavior of this electrodeposition system was gradually dominated by the diffusion process rather than charge transfer process.

5. CIGS thin film by electrodeposition plus periodical mechanical perturbations

The strategy of applying mechanical perturbations to the WE during the electrodeposition process was the result of analyzing the morphology of CIGS film produced by electrodeposition using DC potential at the WE; details of film preparation can be found in [13]. As it was shown with the cross-section micrographs, during the initial stage of CIGS film growth, a more compact CIGS layer is produced. This is evident in the as-electrodeposited film and in the annealed film. In order to promote this formation, a mechanical perturbation to the working electrode was applied every 0.066 C/cm^2 during the electrodeposition processes. With the mechanical perturbation, the solution near the WE, producing perhaps turbulent flow, the diffusion layer tends to disappear for a moment and a new nucleation and growth center was originated, and if the perturbation is periodical, the film will be more compact.

Figure 9(a, b) shows the typical signal of the WE potential and current density versus time collected during the electrodeposition process by applying periodical mechanical perturbations to the WE. The average current density value was 2.4 mA/cm^2. The WE potential and the current density were periodically related to the periodicity of the mechanical perturbations. The WE potential had a variation of −1.0 to −0.995 V at each mechanical perturbation. With the periodical mechanical perturbations, it was possible to make CIGS films with 1.2–1.5 μm in 20 min, and the growth was faster with respect to not using mechanical perturbations. The film composition rations were of Ga/(In+Ga) = 0.28 and Cu/(In+Ga) = 0.93. With the mechanical perturbations, no film dissolution was produced as is presented in pulse reverse electrodeposition.

5.1. Characterization of CIGS films obtained by electrodeposition plus mechanical perturbations

The surface and cross-section morphology of the as-electrodeposited CIGS film are shown in **Figure 10(a, b)**. It is identified that by applying mechanical perturbations, the films are more compact and with less roughness compared to those obtained without applying mechanical

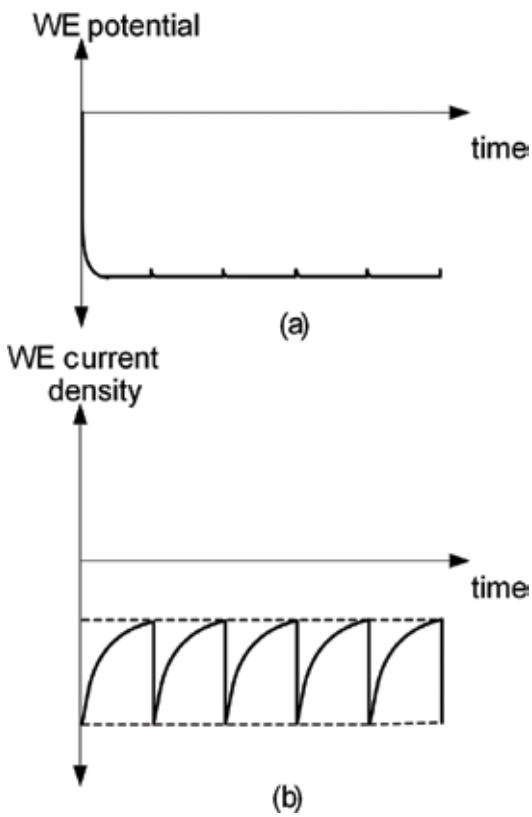

Figure 9. A graphic representation of electrodeposition signals versus time obtained during the electrodeposition process plus periodical mechanical perturbations: (a) WE potential and (b) WE current density [13].

http://dx.doi.org/10.5772/intechopen.78544

perturbations. Films were grown with two different thicknesses, 1200 and 500 nm, that were subjected to the annealing process. The morphology of the annealed films is shown in **Figure 10(c–f)**. In both cases, it is identified that the films are more compact when it is compared to the one obtained by not using mechanical perturbations during the electrodeposition process. As it can be seen on the micrographs, there is coalescence of grains along the film cross section, and the morphology is dense and crack free. Coalescence is achieved due to the fact that the composition is more homogeneous, zones with copper-poor and cooper-rich have been minimized, and the activation energy for grain growth is more uniform throughout the film. This film morphology is completely different from that obtained without applying mechanical perturbations, where there was only coalescence in the first 300 nm of thickness. This is because a more compact

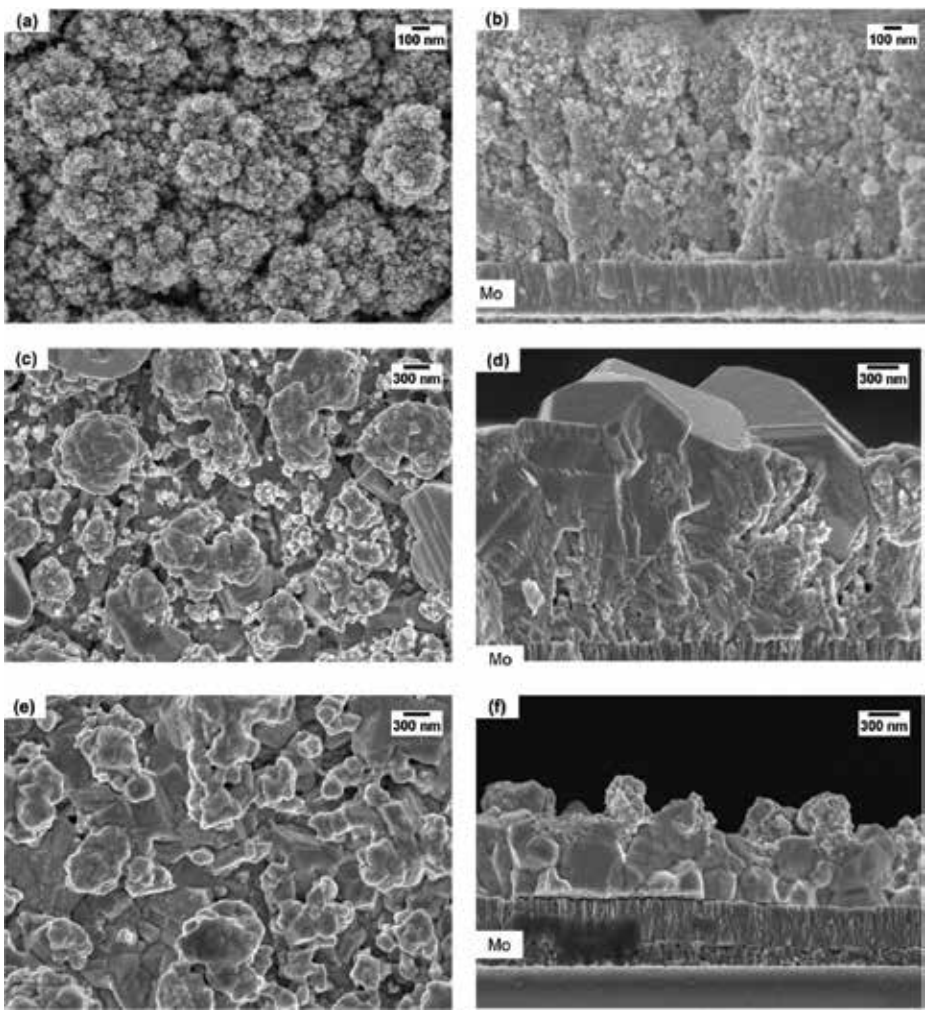

Figure 10. Micrographs of the surface and cross section of the films that have been grown in a potentiostatic mode with mechanical perturbations. (a,b) without annealing, (c,d) annealed films with a thickness of 1200 nm, and (e,f) annealed films with a thickness of 500 nm [13].

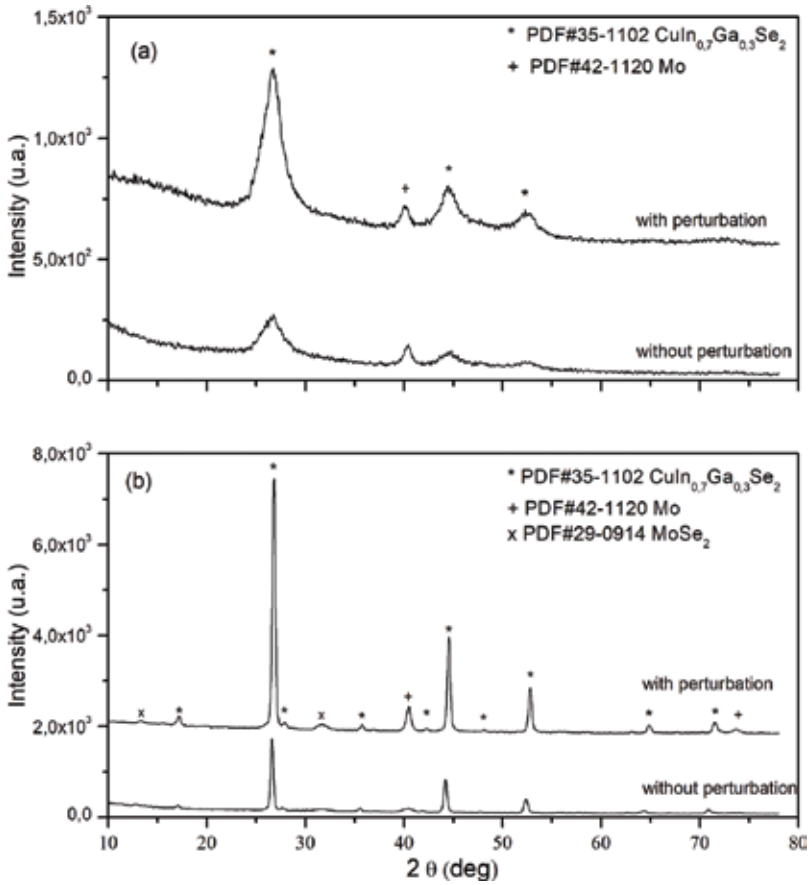

Figure 11. GIXRD diffraction pattern of CIGS films obtained with mechanical perturbation and without mechanical perturbation: (a) as-electrodeposited films and (b) annealed films [13].

morphology was obtained from the electrodeposition process with mechanical perturbations, and it represents a route for obtaining CIGS films by electrodeposition with improved morphology.

Figure 11 shows the GIXRD diffraction pattern for the film deposited with a potentiostatic mode with and without periodical mechanical perturbations with an incidence angle of 1.5°. The $CuIn_xGa_{(1-x)}Se_2$, Mo, and $MoSe_2$ structures are identified according to PDF#35–1102, PDF#42–1120, and PDF#29–0914. First of all, the films exhibit a highly (112) preferred orientation. From **Figure 11(a)**, it can be seen that the as-electrodeposited film shows a poor crystallinity, which is a characteristic of CIGS films before annealing. Also, a diffraction peak of the Mo, which is the substrate and back contact, is identified. The main difference in the diffraction patterns is that there is a greater intensity in the film formed with mechanical perturbation than the film formed without mechanical perturbation. This indicates the presence of higher crystallinity in the film obtained with the mechanical perturbation. **Figure 11(b)** presents the diffraction patterns of the annealed films in selenium atmosphere; these show a high crystalline quality, which is revealed by the well-defined chalcopyrite peaks. The annealing

process clearly increased the grain size, as indicated by the reduction of the peak full-width at half-maximum (FWHM). For the CIGS film formed without mechanical perturbation, the crystal size was 26.6 nm, and for the CIGS film formed with mechanical perturbation, the crystal size was 26.0 nm. This can be expected because the peaks of both films have FWHM that are alike. The annealed films, in a similar manner to the as-electrodeposited film, have a greater intensity in the diffraction peaks for the film formed with mechanical perturbation than for the film formed without mechanical perturbation. Probably, one of the reasons is that the films obtained with mechanical perturbation are denser. The results shown by XRD and SEM clearly demonstrate the advantages of applying periodic perturbations during the electrodeposition process, this being a new route to synthesize thin films.

6. Conclusion(s)

Employing periodic mechanical perturbations during the electrodeposition process allows a better distribution of ionic species on the working electrode surface. This methodology represents a novel approach for the fabrication of thin films by electrodeposition. This has been demonstrated successfully in the synthesis of compact CIGS thin films. It has the advantage to obtain a homogeneous morphology CIGS films in the as-electrodeposited films, as well, in the annealed film. In this strategy, there is no dissolution of the film during the electrodeposition process, as taking place in pulse-reverse electrodeposition. It is a route to obtain CIGS films by electrodeposition with compact morphology and large grains. Further studies should be done about the mechanical perturbation frequency and the effect on the solar cell efficiency.

Acknowledgements

This chapter was supported through the projets C16-FAI-09-58.58, PAPIIT-IN117216, CONACYT-82306 and CONACYT-UNAM (LIFYCS), specially with the use of ICP-AES ULTIMA 2 and SEM S-5500. Also, we would like to thank María Luis Ramón García for the XRD and Rogelio Morán Elvira fo SEM measurements.

Author details

Baudel Lara Lara[1]*, Arturo Fernández Madrigal[2], Lizbeth Morales Salas[2] and Alejandro Altamirano Gutiérrez[3]

*Address all correspondence to: ing_lara@uaslp.mx

1 Facultad de Ingeniería, Universidad Autónoma de San Luis Potosí, Zona Universitaria Poniente, San Luis Potosí, S. L. P., México

2 Instituto de Energías Renovables, Universidad Nacional Autónoma de México, Temixco, Morelos, México

3 Centro Universitario de Tonalá, Universidad de Guadalajara, Tonalá, Jalisco, México

References

[1] Bhattacharya RN. Solution growth and electrodeposited $CuInSe_2$ thin films. Journal of the Electrochemical Society. 1983;**130**:2040-2042. DOI: 10.1149/1.2119516

[2] Peter LM. Electrochemical routes to earth-abundant photovoltaics: A minireview. Electrochemistry Communications. 2015;**50**:88-92

[3] Chen H, Wei Z, Zheng X, Yang S. A scalable electrodeposition route to the low-cost, versatile and controllable fabrication of perovskite solar cells. Nano Energy. 2015;**15**:216-226

[4] Septina W, Ikeda S, Kyoraiseki A, Harada T, Matsumura M. Single-step electrodeposition of a microcrystalline $Cu_2ZnSnSe_4$ thin film with a kesterite structure. Electrochimica Acta. 2013;**88**:436-442

[5] Bhattacharya RN, Oh MK, Kim Y. CIGS-based solar cells prepared from electrodeposited precursor films. Solar Energy Materials and Solar Cells. 2012;**98**:198-202. DOI: 10.1016/j.solmat.2011.10.026

[6] Nady JE, Kashyout AB, Ebrahim S, Soliman MB. Nanoparticles Ni electroplating and black paint for solar collector applications. Alexandria Engineering Journal. 2016;**55**:723-729

[7] Wang J, Lei Z, Zhou Q, Wu W, Zhu C, Liu Z, Chang S, Pu J, Zhang H. Ultra-flexible lithium ion batteries fabricated by electrodeposition and solvothermal synthesis. Electrochimica Acta. 2017;**237**:119-126

[8] Budevski E, Staikov G, Lorenz WJ. Electrocrystallization nucleation and growth phenomena. Electrochimica Acta. 2000;**45**:2559-2574. DOI: 10.1016/S0013-4686(00)00353-4

[9] Taylor SR, Gileadi E. Physical interpretation of the Warburg impedance. Corrosion. 1995;**51**(9):664-671

[10] Fu Y-P, You R-W, Lew K-K. Electrochemical properties of solid–liquid interface of $CuIn_{1-x}Ga_xSe_2$ prepared by electrodeposition with various gallium concentrations. Journal of the Electrochemical Society. 2009;**156**(9):E133-E138. DOI: 10.1149/1.3158558

[11] Zahraei M, Saidi MS, Sani M. Numerical simulation of electro-deposition process influenced by force convection and migration of ions. Journal of Electroanalytical Chemistry. 2016;**782**:117-124. DOI: 10.1016/j.jelechem.2016.012

[12] Scharifker B, Hills G. Theoretical and experimental studies of multiple nucleation. Electrochimica Acta. 1983;**28**(7):879-889. DOI: 10.1016/0013-4686(83)85163-9

[13] Lara-Lara B, Fernández AM. CIGS this film growing by electrodeposition technique using mechanical perturbation at the working electrode. Journal of Materials Science: Materials in Electronics. 2016;**27**:5099-5106. DOI: 10.1007/s10854-016-4400-1

[14] Jackson P, Wuerz R, Hariskos D, Lotter E, Witte W, Powalla M. Effects of heavy alkali elements in cu(in,Ga)Se_2 solar cells with efficiencies up to 22.6%. Physica Status Solidi RRL: Rapid Research Letters. 2016;**10**(8):583-586. DOI: 10.1002/pssr.201600199

[15] Repins I, Contreras MA, Egaas B, DeHart C, Scharf J, Perkins CL, To B, Noufi R. 19·9%-efficient ZnO/CdS/CuInGaSe$_2$ solar cell with 81·2% fill factor. Progress in Photovoltaics: Research and Applications. 2008;**16**:235-239. DOI: 10.1002/pip.822

[16] Saji VS, Choi I-H, Lee C-W. Progress in electrodeposited absorber layer for CuIn$_{(1-x)}$Ga$_x$Se$_2$ (CIGS) solar cells. Solar Energy. 2011;**85**:2666-2678. DOI: 10.1016/j.solener.2011.08.003

[17] Hibberd CJ, Chassaing E, Liu W, Mitzi DB, Lincot D, Tiwari AN. Non-vacuum methods for formation of Cu(In,Ga)(Se,S)$_2$ thin film photovoltaic absorbers. Progress in Photovoltaics: Research and Applications. 2010;**18**:434-452. DOI: 10.1002/pip.914

[18] Lincot D, Guillemoles JF, Taunier S, Guimard D, Sicx-Kurdi J, Chaumont A, Roussel O, Ramdani O, Hubert C, Fauvarque JP, Bodereau N, Parissi L, Panheleux P, Fanouillere P, Naghavi N, Grand PP, Benfarah M, Mogensen P, Kerrec O. Chalcopyrite thin film solar cells by electrodeposition. Solar Energy. 2004;**77**:725-737. DOI: 10.1016/j.solener.2004.05.024

[19] Aksu S, Pinarbasi M. Electrodeposition methods and chemistries for deposition of CIGS precursor thin films. 37th IEEE Photovoltaic Specialists Conference (PVSC). 2011:310-314

[20] Bhattacharya RN, Fernández AM. CuIn$_{1-x}$Ga$_x$Se$_2$-based photovoltaic cells from electrodeposited precursors films. Solar Energy Materials & Solar Cells. 2003;**76**:331-337. DOI: 10.1016/S0927-0248(02)00285-4

[21] Dale P, Peter L. Applications of electrochemistry in the fabrication and characterization of thin-film solar cell. In: Alkire RC, Kolb DM, Lipkowski J, Ross PN, editors. Photoelectrochemical Materials and Energy Conversion Processes. Vol. 12. Wiley-VCH Verlag GmbH Co; 2011. DOI: 10.1002/9783527633227

[22] Roussel O, Ramdani O, Chassaing E, Grand P-P, Lamirand M, Etcheberry A, Kerrec O, Guillemoles J-F, Lincot D. First stages of CuInSe$_2$ electrodeposition from cu(II)-in(III)-se(IV) acidic solutions on polycrystalline Mo films. Journal of the Electrochemical Society. 2008;**155**. DOI: D141–D147. DOI: 10.1149/1.2815476

[23] Chassaing E, Grand P-P, Ramdani O, Vigneron J, Etcheberry A, Lincot D. Electrocrystallization mechanism of cu–in–se compounds for solar cell applications. Journal of the Electrochemical Society. 2010;**157**(7). DOI: D387–D395. DOI: 10.1149/1.3374590

[24] Huang H-C, Lin C-S, Chen F-J, Li W-C. Direct observation of the electrocrystallization of compound CuInSe$_2$ during the early stages of deposition. Electrochimica Acta. 2013;**97**:244-252

[25] Lai Y, Liu J, Yang J, Wang B, Liu F, Zhang Z, Li J, Liu Y. Incorporation mechanism of indium and gallium during electrodeposition of cu(in,Ga)Se$_2$ thin film. Journal of the Electrochemical Society. 2011;**158**. DOI: D704–D709. DOI: 10.1149/2.059112jes

[26] Sang ND, Quang PH, Tu LT, Hop DTB. Effect of electrodeposition potential on composition and morphology of CIGS absorber thin film. Bulletin of Materials Science. 2013;**36**:735-741. DOI: 10.1007/s12034-013-0497-5

[27] Fernández AM, Bhattacharya RN. Electrodeposition of $CuIn_{1-x}Ga_xSe_2$ precursor films: Optimization of film composition and morphology. Thin Solid Films. 2005;**474**:10-13. DOI: 10.1016/j.tsf.2004.02.104

[28] Liu J, Liu F, Lai Y, Zhang Z, Li J, Liu Y. Effects of sodium sulfamate on electrodeposition of cu(in,Ga)Se_2 thin film. Journal of Electroanalytical Chemistry. 2011;**651**:191-196. DOI: 10.1016/j.jelechem.2010.10.021

[29] Calixto ME, Dobson KD, McCandless BE, Birkmire RW. Controlling growth chemistry and morphology of single-bath electrodeposited cu(in,Ga)Se_2 thin films for photovoltaic application. Journal of the Electrochemical Society. 2005;**153**(6). DOI: G521–G528. DOI:10.1149/1.2186764

[30] Fu Y-P, You R-W, Lew KK. $CuIn_{1-x}Ga_xSe_2$ absorber layer fabricated by pulse-reverse electrodeposition technique for thin films solar cell. Journal of the Electrochemical Society. 2009;**156**(12):D553-D557. DOI: 10.1149/1.3240330

[31] Lara-Lara B, Fernández AM. Influence of electrode position in the electrolytic cell configuration for the electrodeposition of cu(in,Ga)Se_2 thin films. Journal of Materials Science: Materials in Electronics. 2015;**26**:5593-5560. DOI: 10.1007/s10854-014-2319-y

[32] Chirilă A, Reinhard P, Pianezzi F, Bloesch P, Uhl AR, Fella C, Kranz L, Keller D, Gretener C, Hagendorfer H, Jaeger D, Erni R, Nishiwaki S, Buecheler S, Tiwari AN. Potassium-induced surface modification of cu(in,Ga)Se_2 thin films for high-efficiency solar cells. Nature Materials. 2013;**12**:1107-1111. DOI: 10.1038/nmat3789

[33] Jackson P, Hariskos D, Lotter E, Paetel S, Wuerz R, Menner R, Wischmann W, Powalla M. New world record efficiency for cu(in,Ga)Se_2 thin-film solar cells beyond 20%. Progress in Photovoltaics: Research and Applications. 2011;**19**:894-897. DOI: 10.1002/pip.1078

[34] Gamburg YD, Zangari G. Theory and Practice of Metal Electrodeposition. New York: Springer; 2011. DOI: 10.1007/978-1-4419-9669-5_1

[35] Kampmann A, Sittinger V, Rechid J, Reineke-Koch R. Large area electrodeposition of cu(in,Ga)Se_2. Thin Solid Films. 2000. DOI: 361-362:309-313. DOI:10.1016/S0040-6090(99)00863-9

[36] Schlenker T, Valero ML, Schock HW, Werner JH. Grain growth studies of thin cu(in, Ga)Se_2 films. Journal of Crystal Growth. 2004;**264**:178-183. DOI: 10.1016/j.jcrysgro.2004.01.020

[37] You R-W, Lew KK, Fu Y-P. Effect of indium concentration on electrochemical properties of electrode-electrolyte interface of $CuIn_{1-x}Ga_xSe_2$ prepared by electrodeposition. Materials Research Bulletin. 2017;**96**:183-187. DOI: 10.1016/j.materresbull.2017.04.027

Chapter 7

Application of the Method of Matched Asymptotic Expansions to Solve a Nonlinear Pseudo-Parabolic Equation: The Saturation Convection-Dispersion Equation

Cíntia Gonçalves Machado and Albert C. Reynolds

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.76828

Abstract

In this work, we apply the method of matched asymptotic expansions to solve the one-dimensional saturation convection-dispersion equation, a nonlinear pseudo-parabolic partial differential equation. This equation is one of the governing equations for two-phase flow in a porous media when including capillary pressure effects, for the specific initial and boundary conditions arising when injecting water in an infinite radial piecewise homogeneous horizontal medium containing oil and water. The method of matched asymptotic expansions combines inner and outer expansions to construct the global solution. In here, the outer expansion corresponds to the solution of the nonlinear first-order hyperbolic equation obtained when the dispersion effects driven by capillary pressure became negligible. This equation has a monotonic flux function with an inflection point, and its weak solution can be found by applying the method of characteristics. The inner expansion corresponds to the shock layer, which is modeled as a traveling wave obtained by a stretching transformation of the partial differential equation. In the transformed domain, the traveling wave solution is solved using regular perturbation theory. By combining the solution for saturation with the so-called Thompson-Reynolds steady-state theory for obtaining the pressure, one can obtain an approximate analytical solution for the wellbore pressure, which can be used as the forward solution which analyzes pressure data by pressure-transient analysis.

Keywords: method of matched asymptotics, boundary layer approximation, nonlinear pseudo-parabolic partial differential equation, convection-dispersion phenomenon, multiphase flow in porous media

1. Introduction

In this chapter, we show how to generate a semi-analytical solution for the wellbore pressure response during a water injection test. In the petroleum industry, well testing is a common practice which consists of wellbore pressure and wellbore flow rate data acquisition in order to estimate parameters that govern flow in the porous media, i.e., the reservoir rock which stores the hydrocarbons. Well tests give an insight into the oil and gas field production potential and profitability and allow the estimation of reservoir parameters. Estimated parameters can be used to calibrate the reservoir numerical simulation model that are used to describe the fluid flow in these reservoirs and forecast their performance as well as to maximize the productivity of the wells. Injections are important tests on reservoirs containing high amount of harmful gases like carbon dioxide and sulfur dissolved in the oil, causing conventional production testing in the exploratory phase of offshore field development inviable. Multiphase flow is the norm in petroleum reservoirs, and an injection test consists of a period of water or gas injection into an oil reservoir (**Figure 1**), a common technique known as waterflooding or gasflooding that is used to displace oil to a producing well. Data from an injection test can be used to estimate the reservoir rock absolute permeability (k), the skin zone permeability (k_s), and the water endpoint relative permeability (a_w). The skin zone permeability is the rock permeability in the zone around the well which was stimulated or damaged during the wellbore drilling operation, while the water endpoint permeability is a measure of how easy water can flow in a specific porous media when there is immobile oil present. In the pursuance of modeling the wellbore pressure response during a water injection test, the Rapport-Leas equation [1], a nonlinear pseudo-parabolic convection-dispersion equation, is used to determine the water saturation distribution in the reservoir as a function of time by assuming a one-dimensional homogeneous medium containing incompressible fluids. Water saturation (S_w) is the fraction of water in a given pore space, and it is expressed in water volume by pore volume. In [2–4], it has been shown how to obtain the wellbore pressure response for the case when capillary

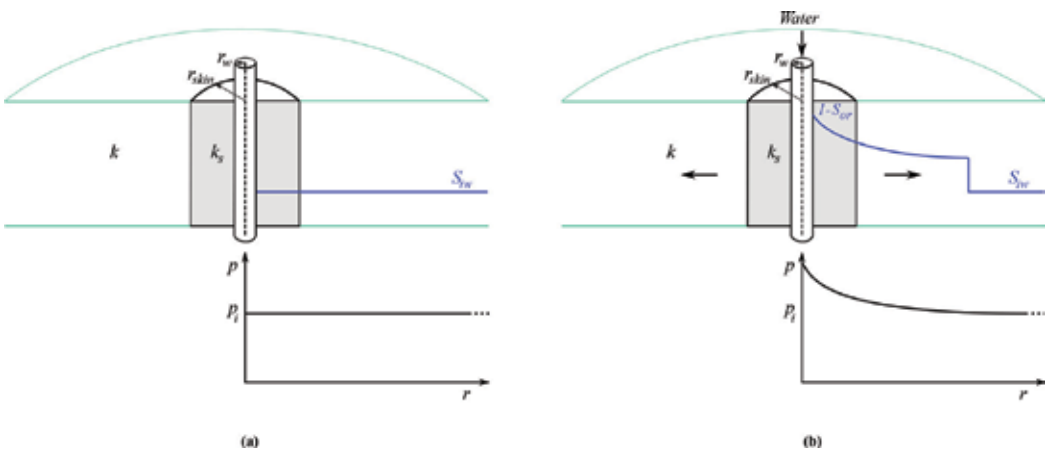

Figure 1. Sketch of the injection test. Reservoir is assumed to be at rest at the beginning of the test with constant pressure and immobile water saturation distribution (a). Water is injected at constant flow rate leading to pressure change that propagates from the well (b).

http://dx.doi.org/10.5772/intechopen.76828

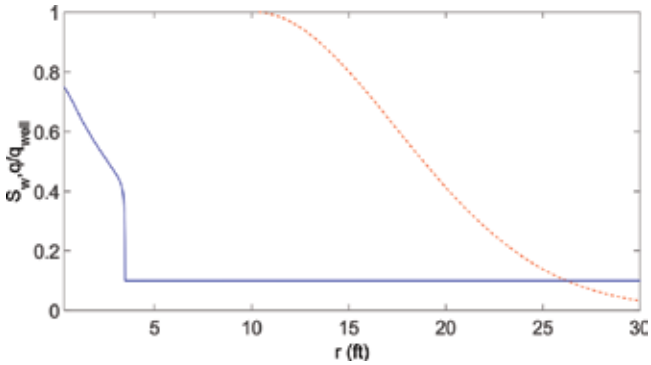

Figure 2. Relationship between the water saturation (S_w), solid blue curve, the dimensionless total flow rate profile (q_D), and dotted red curve, during injection.

pressure effects are negligible, i.e., when dispersion effects are not significant. In this case, the Rapport-Leas equation reduces to the Buckley-Leverett [5] equation, a nonlinear hyperbolic equation. In this work, we have extended their model to include capillary pressure. Although the wellbore pressure during injection seems to be insensitive to capillarity effects insensitive to the accuracy of the calculated saturation distribution in the reservoir, knowledge of the correct saturation profile at the end of injection represents the initial condition and hence is required to calculate the saturation distribution during subsequent tests as shut-in (falloff) and flowback (production) test which would allow the estimation of relative permeabilities and capillary pressure curves. Once the water saturation distribution is determined for each time, the corresponding pressure solution can be obtained by integrating the expression for the pressure gradient, given by Darcy's law, from the wellbore radius to infinity while assuming an infinite-acting reservoir. Because Darcy's law does not assume incompressible flow, the pressure solution is transient and does not need to assume incompressible flow even though the saturation profile is generated from a incompressible assumption. To actually evaluate this integral which represents the pressure solution, however, we must assume that the reservoir rate profile becomes constant in a region from the wellbore to a radius such that all the injected water is contained within the reservoir volume within this radius (**Figure 2**); this radius increases with time [6]. The region within this radius is referred to as the steady-state region or zone. Intuitively, the assumption that this steady-state zone exists appears to be more tenuous as the total compressibility of the system increases. However, the assumption that this steady-state zone exists has shown to yield accurate semi-analytical pressure solutions for gas-condensate systems [7].

2. Mathematical model

The solutions presented assume infinite-acting one-dimensional radial flow from and to a fully penetrating vertical well with no gravity effects. We apply the method of matched asymptotic expansions to solve the one-dimensional saturation convection-dispersion equation, a nonlinear pseudo-parabolic partial differential equation. This equation is one of the governing equations for two-phase flow in a porous media when including capillary pressure effects, for

the specific initial and boundary conditions arising when injecting water in an infinite radial piecewise homogeneous horizontal medium containing oil and water. The method of matched asymptotic expansions combines inner and outer expansions to construct the global solution. In here, the outer expansion corresponds to the solution of the nonlinear first-order hyperbolic equation obtained when the dispersion effects driven by capillary pressure became negligible. This equation has a monotonic flux function with an inflection point, and its weak solution can be found by applying the method of characteristics. The inner expansion corresponds to the shock layer, which is modeled as a traveling wave obtained by a stretching transformation of the partial differential equation. By combining the solution for saturation with the so-called Thompson-Reynolds steady-state theory, one can obtain an approximate analytical solution for the wellbore pressure, which can be used as the forward solution which analyzes pressure data by pressure-transient analysis. Let us start by finding the saturation distribution in the reservoir during injection and show how to find pressure.

2.1. Saturation

The water mass balance equation, in radial coordinates, leads to the following nonlinear partial differential equation [5]:

$$\frac{\partial S_w}{\partial t} + \frac{\theta q_t}{2\pi r h \phi}\frac{\partial F_w(S_w)}{\partial r} = 0, \tag{1}$$

where throughout we assume that porosity (ϕ) is homogeneous; q_t is the total liquid rate in RB/D; θ represents in general a unit conversion factor where in the oil field units used here, θ =5.6146/24; the reservoir thickness, h, and the radius, r, are in ft; and time, t, is in hours. Let us use Darcy's equation in radial coordinates without gravity for the oil (o) and water (w) flow rate in RB/D given by

$$q_p = -\frac{k(r)h\lambda_p(S_w)}{\alpha}\left(r\frac{\partial p_p}{\partial r}\right), \qquad \text{for } p = o, w. \tag{2}$$

For field units used throughout, α = 141.2. p_p, is the phase p pressure. The λ_p is the phase p mobility, given by the ratio of the phase permeability (k_{rw} or k_{ro}), which are functions of the water saturation, by the phase viscosity (μ_w or μ_o). To find the water fractional flow (F_w), we can subtract Eq. (2) for water from Eq. (2) for oil, to get

$$\frac{\alpha q_o}{k(r)h\lambda_o(S_w)} - \frac{\alpha q_w}{k(r)h\lambda_w(S_w)} = -\left(r\frac{\partial p_o}{\partial r} - r\frac{\partial p_w}{\partial r}\right). \tag{3}$$

Rearranging Eq. (3), substituting the capillary pressure p_c given by the difference of the oil pressure (p_o) and the water pressure (p_w), and dividing the resulting equation by the total flow rate, q_t, yield

$$F_o - \frac{F_w\lambda_o(S_w)}{\lambda_w(S_w)} = -\frac{k(r)h\lambda_o(S_w)}{\alpha q_t}\left(r\frac{\partial p_c(S_w)}{\partial r}\right), \tag{4}$$

http://dx.doi.org/10.5772/intechopen.76828

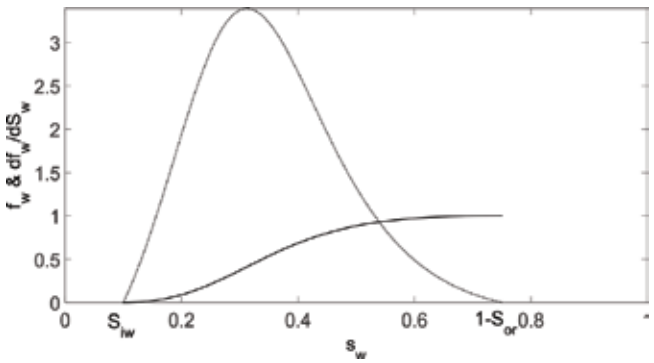

Figure 3. Water fractional flow curve (dark solid curve) and its derivative (dotted curve).

where F_o and F_w are the oil and water fractional flow given by q_o and q_w, respectively. We assume throughout that water is the wetting phase. Finally, substituting $F_o = 1 - F_w$ in Eq. (4) and solving for F_w, we have the following expression for the water fractional flow including capillary pressure effects

$$F_w(S_w) = \frac{1 + \frac{k(r)h\lambda_o(S_w)}{\alpha q_t}\left(r\frac{\partial p_c}{\partial r}\right)}{1 + \frac{\lambda_o(S_w)}{\lambda_w(S_w)}} = \frac{1}{1 + \frac{\lambda_o(S_w)}{\lambda_w(S_w)}} + \frac{\frac{k(r)hk_{ro}}{\alpha q_t \mu_o}\left(r\frac{\partial p_c}{\partial r}\right)}{1 + \frac{\lambda_o(S_w)}{\lambda_w(S_w)}} = f_w + \epsilon r k(r) f_w(S_w) k_{ro}(S_w)\frac{\partial p_c(S_w)}{\partial r}, \tag{5}$$

where f_w is the water mobility ratio (**Figure 3**), i.e., the ratio of water mobility and the total mobility (λ_t), given by

$$f_w(S_w) = \frac{1}{1 + \frac{\lambda_o(S_w)}{\lambda_w(S_w)}} = \frac{\lambda_w(S_w)}{\lambda_o(S_w) + \lambda_w(S_w)}\frac{\lambda_w(S_w)}{\lambda_o(S_w) + \lambda_w(S_w)}, \tag{6}$$

which usually assumes an S-shape. ϵ is the perturbation parameter, defined by

$$\epsilon = \frac{h}{\alpha q_t \mu_o} \tag{7}$$

and the permeability is a function of radius because we consider a skin-damaged zone:

$$k(r) = \begin{cases} k_s, & r_w \le r < r_{\text{skin}} \\ k, & r \ge r_{\text{skin}}, \end{cases} \tag{8}$$

where r_w is the wellbore radius, r_{skin} is the radius of the damaged zone, and k_s is the permeability in the skin zone. Grouping all the parameters that are the function of water saturation, we can define (**Figure 4**)

$$\frac{d\Psi}{dS_w}(S_w) = -f_w(S_w)k_{ro}(S_w)\frac{dp_c}{dS_w}(S_w), \tag{9}$$

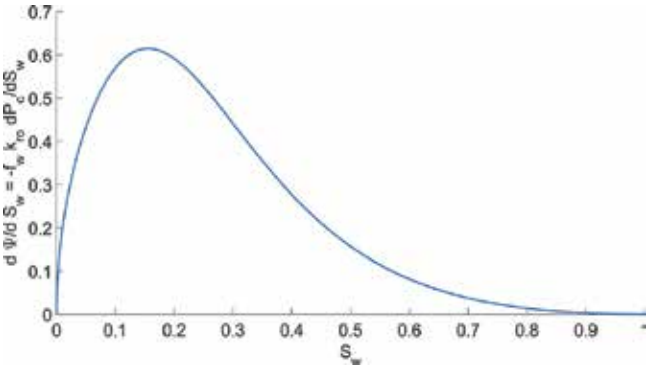

Figure 4. $\frac{d\Psi}{dS_w}$ versus water saturation for a partially water wet reservoir (b). For a strong water, we reservoir, $\frac{d\Psi}{dS_w} \to \infty$ at $S_w = S_{iw}$.

and rewrite Eq. (10) as

$$F_w(S_w) = f_w(S_w) - \epsilon r k(r) \frac{d\Psi(S_w)}{dS_w} \frac{\partial S_w}{\partial r}. \tag{10}$$

For simplicity, we use the Brooks and Corey model [8] given by

$$p_c(S_w) = p_t \left(sv + \frac{S_w - S_{wi}}{1 - S_{wi} - S_{or}} \right)^{-\frac{1}{\lambda}}, \tag{11}$$

to represent capillary pressure. Here, S_{iw} is the immobile water saturation and S_{or} is the residual oil saturation. λ, where $0.4 \leq \lambda \leq 4.0$, is a measure of the pore size distribution (the greater the λ value, the more uniform is the pore size distribution), and p_t is the threshold pressure. The threshold pressure is a measure of the maximum pore size [9], i.e., the minimum capillary pressure at which a continuous nonwetting phase exists in the imbibition case and a continuous wetting phase exists in the drainage case [10]. The greater is the maximum pore size, the smaller is the pressure threshold. According to [11], the extrapolation of the capillary pressure curve obtained from experimental data to $S_w = 1$ yields the correct threshold value. In practice, we introduce a small variable, sv, to limit the maximum value of p_c to a finite value. We can relate the relative permeabilities and the capillary pressure through λ by using the [12] model for the water phase (wetting phase)

$$k_{rw} = a_w \left(\frac{S_w - S_{iw}}{1 - S_{iw} - S_{or}} \right)^{\frac{2+\lambda}{\lambda}}, \tag{12}$$

and the [8] model for the oil phase (nonwetting phase) [13]

$$k_{ro} = \left(1 - \frac{S_w - S_{iw}}{1 - S_{iw} - S_{or}} \right)^2 \left(1 - \left(\frac{S_w - S_{iw}}{1 - S_{iw} - S_{or}} \right)^{\frac{2+\lambda}{\lambda}} \right). \tag{13}$$

http://dx.doi.org/10.5772/intechopen.76828

Now that we have defined the fractional flow rate and its parameters, let us go back to our governing equation for saturation (Eq. (1)). Inserting Eq. (10) into Eq. (1) and defining

$$C = \frac{\theta q_t}{\pi h \phi}, \tag{14}$$

yields

$$\frac{\partial S_w}{\partial t} + \frac{C}{2r}\frac{\partial f_w}{\partial r} - \epsilon \frac{C}{2r}\frac{\partial}{\partial r}\left(rk(r)\frac{\partial \Psi}{\partial r}\right) = 0, \tag{15}$$

which is the nonlinear "pseudo-parabolic" governing equation for saturation. If we insert same common values for the parameters in Eq. (7) to have an idea of its order of magnitude, we can see that epsilon is a very small number. This suggests that the effect of the third term in Eq. (15) may be treated as a perturbation to the first-order hyperbolic equation [5], given by

$$\frac{\partial S_w}{\partial t} + \frac{C}{2r}\frac{\partial f_w}{\partial r} = 0, \tag{16}$$

where f_w is considered to be an S-shaped function along this chapter. During injection, for a partially water wet reservoir, the capillary pressure dispersive effect will be non-negligible only in a small region around the water front (hypodispersion phenomenon) [14, 15] where the capillary pressure derivative and the saturation gradient are significant (**Figure 5**). The capillary pressure smears the water front during injection balancing the self-sharpening tendency of the shock. [16] have developed an exact analytical solution for linear waterflood including the effects of capillary pressure, but their solution is limited to a particular functional form to represent relative permeabilities and capillary pressure curves and does not consider radial flow, which makes their solution very restrictive. As done by [17, 18] for Cartesian coordinates and by [19] for streamlines and streamtubes, the perturbation caused by the capillary pressure effects can be modeled as a shock layer (water front) which moves with the same speed as the shock wave. By applying the method of matched asymptotic expansions [20, 21], we can

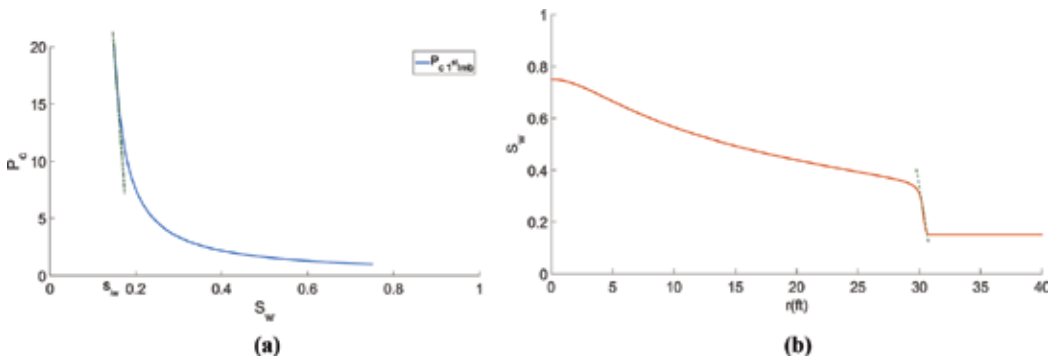

Figure 5. Capillary pressure curve (a) and saturation profile at a time t during water injection (b). The dashed green lines represent the capillary pressure derivative (a) and the saturation gradient (b) at the water front.

combine the solution of the Buckley-Leverett equation (Eq. (16)) with this steady traveling wave to generate an approximate solution of the Rappaport and Leas equation, i.e., the solution of the convection–dispersion saturation equation. In order to solve the [1] equation for the injection period, with the following initial and boundary conditions

$$S_w(r,0) = S_{iw}, \tag{17}$$

$$F_w(r_w,t) = 1, \tag{18}$$

$$\lim_{r\to\infty} S_w = S_{iw}. \tag{19}$$

we divide the domain into two regions, outer and inner regions (**Figure 6**), where the inner region, the region around the water front, is modeled as a shock layer which propagates with the same speed as the shock would be obtained when $\epsilon \to 0$, i.e., when the capillary pressure effects are null. The combination of the self-sharpening tendency of the shock ($S_{wf} > S_{iw}$) with the dispersive effect from the capillary pressure balance against each other leads to the shock layer [22]. Note: in order to guarantee pressure continuity, we have assumed that the capillary pressure gradient is zero at the wellbore, which means that $F_w = f_w = 1$ is the wellbore, so $S_w(r_w,t) = 1 - S_{or}$. This boundary condition will be used for both the Buckley-Leverett and the Rapoport-Leas solutions. Ref. [23] presented the idea of using the method of matched asymptotic equations to solve the Rapoport-Leas equation, while [18, 24, 19] showed how the mass balance could be used to present a closed solution for the saturation distribution. Ref. [24] derived an approximate solution for the [1] in Cartesian coordinates for both water and oil injections into a core considering end effects by also applying the method of matched asymptotic expansions. The method of asymptotic expansions uses the inner and outer saturation solutions combined with a matching function in order to obtain a composite solution which avoids abruptly switching from the outer to the inner solution or vice versa. The inner and outer solutions are each capable of representing the real solution in two distinct regions—the "inner region" and the "outer region" of the boundary layer (**Figure 6**). Similarly, we approximate the saturation solution of Rapoport-Leas equation by forming a composite solution

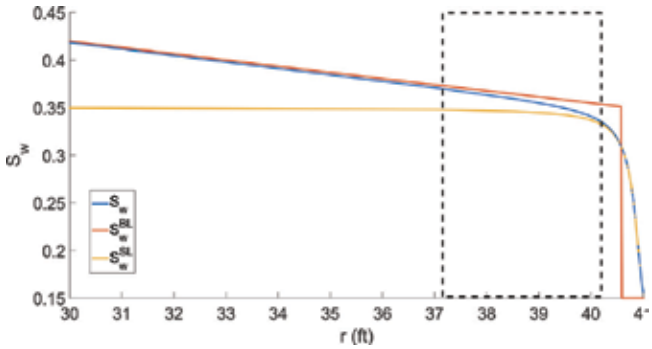

Figure 6. True saturation distribution in the reservoir (S_w) compared with the outer solution (S_w^{BL}) and the inner solution (S_w^{SL}). The dashed square shows the saturation transition zone between the outer and the inner solution where none of these two are capable of approximate S_w.

http://dx.doi.org/10.5772/intechopen.76828

given by the combination of three saturations: S_w^{BL}, the solution obtained when the capillary pressure effects are neglected; S_w^{SL}, the saturation distribution in the shock layer obtained by magnifying the dispersion effects in the saturation governing equation; and S_w^{SH}, the shock wave represented by a Heaviside function:

$$S_w(r,t) \simeq S_w^{BL}(r,t) + S_w^{SL}(r,t) - S_w^{SH}(r,t), \tag{20}$$

where *BL* stands for Buckley-Leverett, *SL* for shock layer, and *SH* for shock function.

2.1.1. Outer solution (S_w^{BL})

The outer solution, S_w^{BL}, is obtained by letting $\epsilon \to 0$ in Eq. (15):

$$S_w^{BL}(r,t) = \lim_{\epsilon \to 0,\ (r,t)\ \text{fixed}} S_w(r,t,\epsilon). \tag{21}$$

That is the nonlinear hyperbolic convection equation known as the Buckley-Leverett saturation equation given by Eq. (16) which is obtained when capillary and gravity effects are neglected. The well-known unique admissible weak solution of this Riemann problem, with the following initial condition

$$S_w^{BL}(r,0) = \begin{cases} 1 - S_{or}, & \text{for } r \leq r_w \\ S_{iw}, & \text{for } r > r_w, \end{cases} \tag{22}$$

can be obtained by the application of the method of characteristics and is given by [5].

$$S_w^{BL}(r,t) = \begin{cases} 1 - S_{or}, & r^2 \leqslant r_w^2 \\ \left(\frac{df_w}{dS_w}\right)^{-1}\left(\frac{1}{C}\frac{\left(r^2 - r_w^2\right)}{t}\right), & r_w^2 < r^2 \leqslant r_w^2 + Dt \\ S_{iw}, & r^2 > Dt + r_w^2, \end{cases} \tag{23}$$

that is, by a family of rarefaction waves, a semi-shock wave, and a constant saturation zone where water is immobile. The shock jump is caused by the S-shaped form of the fractional flow curve, which leads to a gradient catastrophe and consequently a shock solution. This semi-shock has a constant speed, satisfying the Rankine-Hugoniot condition [25]:

$$D = C\frac{\left[f_w\left(S_{wf}\right) - f_w(S_{iw})\right]}{\left[S_{wf} - S_{iw}\right]}, \tag{24}$$

where S_{wf} and S_{iw} are the shock saturations. In this case, in order to satisfy the conservation of mass, the shock speed should correspond to the slope of a tangent line to the water fractional flow curve, i.e.,

$$\frac{f_w\left(S_{wf}\right) - f_w(S_{iw})}{S_{wf} - S_{iw}} = \frac{df_w\left(S_{wf}\right)}{dS_w}. \tag{25}$$

The details of this solution can be found in [5]. **Figure 7** shows the shock jump slope tangent to the fractional flow curve at $S_w = S_{wf}$ and the saturation distribution in the reservoir at a time t. The rarefaction wave family spans from $1 - S_{or}$ to S_{wf} from r_w to $r = 25$ ft, the water front position, i.e., the shock front position, r_s. Ahead of the water front position, there is an immobile water. **Figure 8** compares this solution, the outer solution, with the true solution; there is the convection-dispersion saturation profile. Here, we call the true solution the solution obtained from a numerical simulator.

2.1.2. *Inner solution* (S_w^{SL})

As mentioned, the inner solution intends to represent the saturation profile in the "inner" region around the water front, which is a shock layer (a boundary layer) around the shock traveling

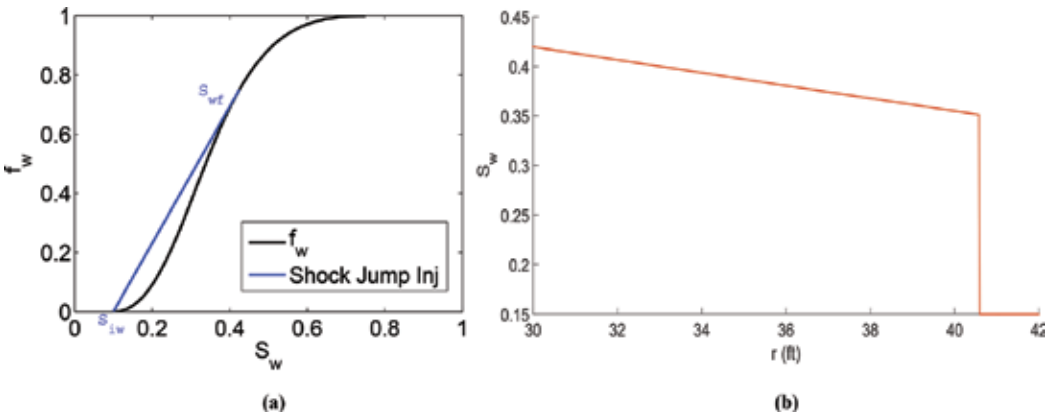

Figure 7. The shock jump slope tangent (blue curve) to the S-shaped fractional flow curve at $S_w = S_{wf}$ (a) and the saturation profile in the reservoir at a time t (b). The rarefaction waves family spans from $1 - S_{or}$ 1 to S_{wf} and from r_w to r = 25 ft, the water front position, i.e., the shock front position, $r_{f,inj}$. Ahead of the water front position, there is immobile water.

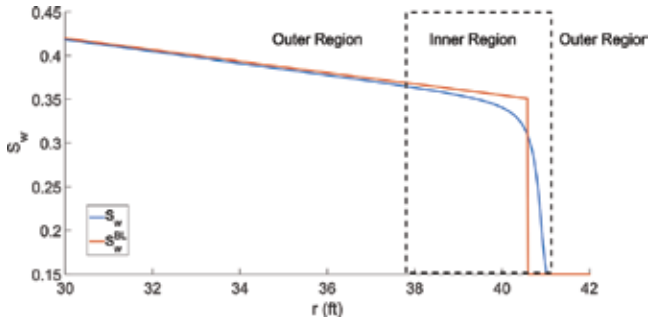

Figure 8. Saturation distribution during the injection period for the outer solution (S_w^{BL}), without capillary pressure, and for the true solution (S_w), with capillary pressure. Both profiles agree in the region far from the water front, the region outside the dashed square.

http://dx.doi.org/10.5772/intechopen.76828

with the same speed as the shock itself (**Figure 9**). In order to find S_w^{SL}, we magnify the shock layer by using a stretching traveling wave coordinate. Similarly, as defined in [24, 18, 19]

$$w = w(r,t) = \frac{r^2 - r_s^2(t)}{\epsilon}, \tag{26}$$

where r_s is the shock front position:

$$r_s^2(t) = r_w^2 + CDt, \tag{27}$$

w is zero at $r = r_s$ and goes to $\pm\infty$ as $\epsilon \to 0$. We rewrite Eq. (15) in terms of moving coordinates, $(r,t) \to (w,\tau)$, where $\tau = \tau(t) = t$. Using Eq. (26) in the transformed equation and multiplying the resulting equation by ϵ yield

$$\frac{\epsilon}{C}\frac{\partial S_w}{\partial \tau} - D\frac{\partial S_w}{\partial w} + \frac{\partial f_w}{\partial w} - \frac{\partial}{\partial w}\left(2\big(\epsilon w + r_s^2(\tau)\big)k\big(\epsilon w + r_s^2(\tau)\big)\frac{\partial \Psi}{\partial w}\right) = 0. \tag{28}$$

The inner solution is obtained by letting $\epsilon \to 0$ in Eq. (28):

$$S_w^{SL}(w,\tau) = \lim_{\epsilon \to 0,\ (w,\tau)\ \text{fixed}} S_w\big(\epsilon w + r_s^2(\tau), \tau, \epsilon\big), \tag{29}$$

as presented in [26]. Therefore, neglecting the terms of order ϵ in Eq. (28), we have

$$-D\frac{\partial S_w^{SL}}{\partial w} + \frac{\partial f_w}{\partial w} - \frac{\partial}{\partial w}\left(2r_s^2(\tau)k\big(r_s^2(\tau)\big)\frac{\partial \Psi}{\partial w}\right) = 0. \tag{30}$$

Note that here we are treating the permeability k as function of the shock position radius, r_s, only, by assuming that in the limit of the inner solution, $\epsilon(r \to r_s(\tau))$. Intuitively, this assumption does not seem valid when the shock layer is crossing heterogeneity interfaces, i.e., interfaces between two different permeability zones. However, for the water injection in a field

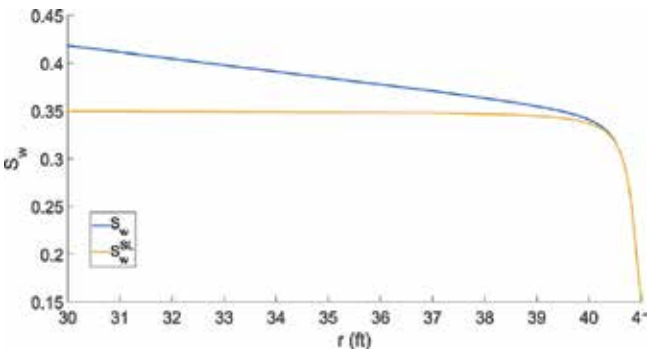

Figure 9. Saturation distribution during the injection period for the inner solution (S_w^{SL}) and for the true solution (S_w), with capillary pressure. Both profiles agree in the region around the water front, i.e., in the shock boundary layer which we have defined as the inner region.

scale, the skin zone will be crossed by the water front in a very short time, and we will only need to use the pseudo-parabolic equation (Eq. (15)) to find saturation for the end of injection period (to be used as initial condition for falloff and flowback tests, as mentioned in the introduction). Consequently, we can simplify the problem as shown above. Integrating the ordinary differential equation given by Eq. (30) with respect to w for any fixed time τ and applying the chain rule gives

$$-DS_w^{SL} + f_w\left(S_w^{SL}\right) - 2r_s^2(\tau)k\left(r_s^2(\tau)\right)\frac{d\Psi\left(S_w^{SL}\right)}{dS_w}\frac{\partial S_w^{SL}(w,\tau)}{\partial w} = a(\tau), \tag{31}$$

where $a(\tau)$ is constant for the injection case, as we will show later. As mentioned, the inner solution is modeled as a traveling wave with a constant speed—the shock speed—and the boundary conditions (for the inner solution) given by

$$\begin{cases} w \to \infty: & S_w^{SL} = S_{iw}, \quad \dfrac{\partial S_w^{SL}}{\partial w} = 0, \\ w \to -\infty: & S_w^{SL} = S_{wf}, \quad \dfrac{\partial S_w^{SL}}{\partial w} = 0, \end{cases} \tag{32}$$

as the inner solution goes asymptotically to the shock saturations. This necessity of this behavior will be clearer very soon when we compare the inner solution with the matching saturation solution. Using the first boundary condition given by Eq. (32) in Eq. (31) leads to

$$a(\tau) = -DS_{iw}, \tag{33}$$

while using the second boundary condition given by Eq. (32) yields

$$a(\tau) = -DS_{wf} + f_w\left(S_{wf}\right); \tag{34}$$

implying that $-D\ (\ S\_\{iw\} - S\_\{wf\}) - f\_w(S\_\{wf\}) = 0$, which it is indeed correct from the definition of D in Eq. (24). As we can see from Eqs. (33) and (34), $a(\tau)$ is a constant and it will be called simply by a from now on. Substituting the constant a (Eq. (33)) in Eq. (31) and dividing it by $D\left(S_{iw} - S_w^{SL}\right) + f_w\left(S_w^{SL}\right)$ yield

$$2r_s^2(\tau)k\left(r_s^2(\tau)\right)\frac{\frac{d\Psi\left(S_w^{SL}\right)}{dS_w}}{D\left(S_{iw} - S_w^{SL}\right) + f_w\left(S_w^{SL}\right)}\frac{\partial S_w^{SL}(w,\tau)}{\partial w} = 1. \tag{35}$$

Integrating Eq. (35) from $w_{well} = w(r_w, \tau)$ to any w at any time τ gives us the relationship between any S_w^{SL} and w:

$$2r_s^2 k\left(r_s^2\right)\int_{S_w^{SL}(w_{well})}^{S_w^{SL}} \frac{\frac{d\Psi\left(S_w^{SL}\right)}{dS_w}}{D\left(S_{iw} - S_w^{SL}\right) + f_w\left(S_w^{SL}\right)} dS_w^{SL} = \int_{w_{well}}^{w} dw. \tag{36}$$

At $S_w^{SL} = S_{wf}$, the integral in the left side of Eq. (36) diverges as the integrand denominator goes to 0. This behavior is consistent with our boundary condition assumptions for S_w^{SL} (Eq. (32)). At

$S_w^{SL} = S_{iw}$, the integral in the left side of Eq. (36) does converge when the integrand numerator also goes to zero (**Figure 4**), a behavior which is consistent when trying to model a hypodispersion phenomenon (for a partially water wet reservoir). In another reservoir wettability scenario, e.g., a strong water wet rock, the capillary pressure would not be bounded at Siw, and the integral in the left side of Eq. (36) would diverge. Note: we still do not know the value of S_w^{SL} at w well. In order to find a closed form for this problem, mass balance can be used, but first let us present the matching saturation, since this solution will be necessary for the mass balance.

2.1.3. Matching solution (S_w^{SH})

The matching saturation S_w^{SH} is defined using the matching principle by applying Prandtl's technique [26]:

$$\lim_{r^2 \to r_w^2 + CDt} S_w^{BL}(r,t) = \lim_{w \to \pm\infty} S_w^{SL}(w,t); \tag{37}$$

and in the injection case is given by

$$S_w^{SH}(r,t) = \begin{cases} S_{iw}, & r^2 \geq r_s^2(t) = r_w^2 + CDt, \\ S_{wf}, & r_w^2 \leq r_s^2(t). \end{cases} \tag{38}$$

which is plotted in **Figure 10** against the outer and inner solutions. As we were searching for, S_w^{SH} matches with the outer solution in the inner region and with the inner solution in the outer region, being able to subtract their effect in the composite solution in their "non-correspondent" zones. **Figure 11** compares the saturation distribution during the injection period for the true solution obtained from the numerical simulator IMEX with the outer, inner, and matching saturation solutions.

2.1.4. Mass balance

Now that we have defined all the three saturations that are composed of the approximate solution for the convection dispersion saturation equation, let us try to find a closed form for

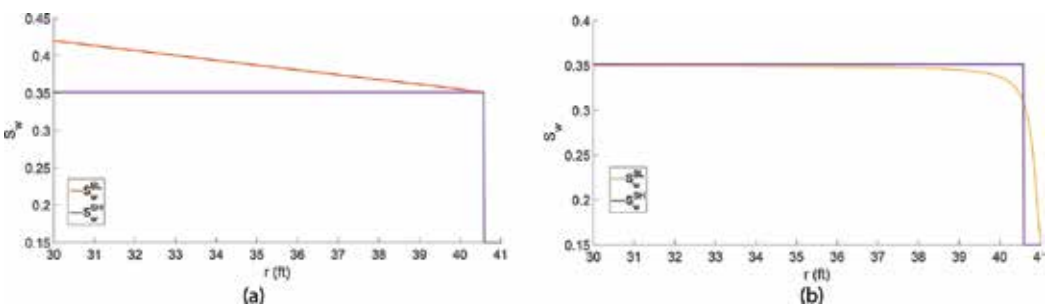

Figure 10. The matching saturation function (S_w^{SH}) compared with the outer solution (S_w^{BL}) (a) and with the inner solution (S_w^{SH}) (b). S_w^{SH} matches with the outer solution in the inner region and with the inner solution in the outer region, as desired.

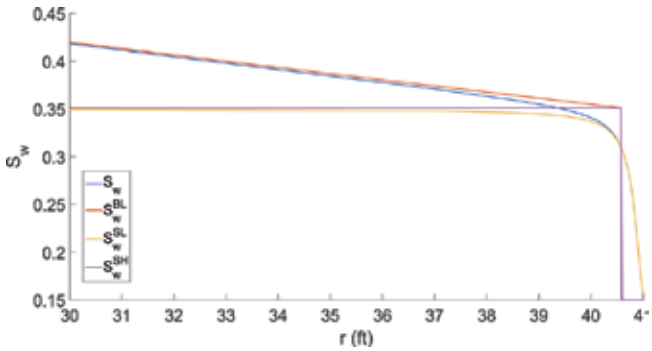

Figure 11. Saturation distribution during the injection period with (true solution) and without capillary pressure (outer solution), the traveling wave (inner solution), and the matching saturation (a). S_w^{SH} matches with the region inner solution in the outer region, the region far from the water front.

the saturation distribution based in the mass balance. Since both the Buckley-Leverett (S_w^{BL}) solution and the composite solution (S_w) must obey material balance, the following equality

$$q_t t = \int_{r_w^2}^{\infty} (S_w(r,t) - S_{iw})\pi h dr^2 = \int_{r_w^2}^{\infty} \left(S_w^{BL}(r,t) - S_{iw}\right)\pi h dr^2. \tag{39}$$

must hold. From Eq. (20) and Eq. (39), it follows that

$$\int_{r_w^2}^{\infty} \left(S_w^{BL} + S_w^{SL} - S_w^{SH} - S_{iw}\right)\pi h dr^2 = \int_{r_w^2}^{\infty} \left(S_w^{BL}(r,t) - S_{iw}\right)\pi h dr^2, \tag{40}$$

which, upon simplification, gives

$$\int_{r_w^2}^{\infty} \left(S_w^{SL} - S_w^{SH}\right) dr^2 = 0. \tag{41}$$

Rearranging Eq. (41) using Eq. (38) for S_w^{SH} gives

$$\int_{r_w^2}^{\infty} S_w^{SL} dr^2 = \int_{r_w^2}^{r_s^2} S_{wf} dr^2 + \int_{r_s^2}^{\infty} S_{iw} dr^2 = S_{wf}\left(r_s^2 - r_w^2\right) + \left(S_{iw}\int_{r_w^2}^{\infty} dr^2 - S_{iw}\int_{r_w^2}^{r_s^2} dr^2\right). \tag{42}$$

Using Eq. (27) in Eq. (42), it follows that

$$\int_{r_w^2}^{\infty} S_w^{SL} dr^2 = \left(S_{wf} - S_{iw}\right) CDt + S_{iw}\int_{r_w^2}^{\infty} dr^2. \tag{43}$$

Transforming Eq. (43) from $(r,t) \rightarrow (w,\tau)$ and using Eq. (26), Eq. (43) becomes

$$\epsilon \int_{-\frac{CD\tau}{\epsilon}}^{\infty} S_w^{SL}(w) dw = \left(S_{wf} - S_{iw}\right) CD\tau + \epsilon S_{iw}\int_{-\frac{CD\tau}{\epsilon}}^{\infty} dw. \tag{44}$$

From Eq. (35),

$$dw = 2r_s^2 k(r_s^2)\frac{\frac{d\Psi}{dS_w}}{D(S_{iw} - S_w^{SL}) + f_w}dS_w^{SL}. \tag{45}$$

Substituting Eq. (45) in Eq. (44) and solving the resulting equation divided by ϵS_{iw} for

$$\int_{-\frac{CD\tau}{\epsilon}}^{\infty} dw = \frac{2r_s^2 k(r_s^2)}{S_{iw}}\int_{S_w^{SL}\left(-\frac{CD\tau}{\epsilon}\right)}^{S_{iw}} S_w^{SL}\frac{\frac{d\Psi}{dS_w}(S_w^{SL})}{D(S_{iw} - S_w^{SL}) + f_w(S_w^{SL})}dS_w^{SL} - \frac{(S_{wf} - S_{iw})CD\tau}{\epsilon S_{iw}}. \tag{46}$$

Setting $S_w = S_{iw}$ in the upper limits of the integrals of Eq. (36) and exchanging the two sides of the equation yield

$$\int_{-\frac{CD\tau}{\epsilon}}^{\infty} dw = 2r_s^2 k(r_s^2)\int_{S_w^{SL}\left(-\frac{CD\tau}{\epsilon}\right)}^{S_{iw}} \frac{\frac{d\Psi}{dS_w}(S_w^{SL})}{D(S_{iw} - S_w^{SL}) + f_w(S_w^{SL})}dS_w^{SL}. \tag{47}$$

As the left sides of Eqs. (46) and (47) are the same, the right sides of these two equations must be equal which gives

$$2r_s^2 k(r_s^2)\int_{S_w^{SL}\left(-\frac{CD\tau}{\epsilon}\right)}^{S_{iw}} \frac{\frac{d\Psi}{dS_w}(S_w^{SL})dS_w^{SL}}{D(S_{iw} - S_w^{SL}) + f_w(S_w^{SL})} = \frac{2r_s^2 k(r_s^2)}{S_{iw}}\int_{S_w^{SL}\left(-\frac{CD\tau}{\epsilon}\right)}^{S_{iw}} \frac{\frac{S_w^{SL}d\Psi}{dS_w}(S_w^{SL})dS_w^{SL}}{D(S_{iw} - S_w^{SL}) + f_w(S_w^{SL})} - \frac{(S_{wf} - S_{iw})CD\tau}{\epsilon S_{iw}}. \tag{48}$$

Multiplying Eq. (48) by ϵS_{iw} and rearranging the resulting equation give

$$2r_s^2 k(r_s^2)\epsilon\int_{S_w^{SL}\left(-\frac{CD\tau}{\epsilon}\right)}^{S_{iw}} (S_w^{SL} - S_{iw})\frac{\frac{d\Psi}{dS_w}(S_w^{SL})}{D(S_{iw} - S_w^{SL}) + f_w(S_w^{SL})}dS_w^{SL} = (S_{wf} - S_{iw})CD\tau. \tag{49}$$

Once the value $S_w^{SL}\left(-\frac{CD\tau}{\epsilon}\right)$ (i.e., the inner solution saturation in the wellbore $S_w^{SL}(w_{well})$) is determined numerically by solving Eq. (49) using the bisection method at each time τ, Eq. (36) is used to determine the saturation profile in the stabilized zone. It is important to note that, as S_w^{SL} should reach S_{wf} and S_{iw} asymptotically as $w \to \pm\infty$, here we did not have to fix a finite distance in which the traveling wave would reach its open bounds as done by [18, 19, 24]. With our approach, as shown in the validation section (**Figure 12**), we can obtain essentially a perfect match with the numerical solution, with a "smoother" water front, which is expected from the dispersive effect of capillary pressure, contrary to the sharp transition between the saturation at the water front foot (r_{wf}) which is the finite position at which water can be considered immobile—and the initial water saturation in the oil zone exhibited by solutions of previous authors.

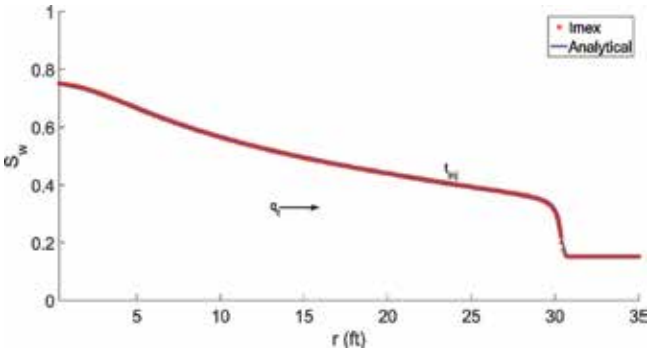

Figure 12. Comparison of the saturation distribution from analytical solution and IMEX during the injection period with capillary pressure.

2.2. Wellbore pressure

As mentioned previously, after finding the saturation distribution, we can obtain the wellbore pressure by applying the pressure solutions presented by [2]. During injection at a constant flow rate, $q_t(r_w, t)$ RB/D, where $t = 0$ at the beginning of the water injection, by integrating Darcy's law as in [6, 2], given by

$$q_t = -\frac{k(r)hr}{\alpha}\left(\lambda_t \frac{\partial p_o}{\partial r} - \lambda_w \frac{\partial p_c}{\partial r}\right), \tag{50}$$

where $p_w = p_o - p_c$. Eq. (50) can be solved for the oil pressure gradient by integrating it from the wellbore radius to infinite, assuming an infinite-acting reservoir. The bottom hole pressure difference from the reservoir initial pressure (p_{oi}) can then be expressed as

$$\Delta p_{wf}(t) = p_{wf}(t) - p_{oi} = \int_{r_w}^{\infty} \frac{\alpha q_t(r,t)}{h\lambda_t(r,t)k(r)} \frac{dr}{r} - \int_{r_w}^{\infty} f_w \frac{dp_c}{dS_w}\frac{\partial S_w}{\partial r}\, dr, \tag{51}$$

where it is assumed that $p_o(r_w, t) = p_w(r_w, t)$, i.e., $p_c = 0$ at $r = r_w$, in order to satisfy the compatibility condition [27], i.e., to guarantee phase pressure continuity at the wellbore. Eq. (51) can be rewritten as

$$\Delta p_{wf}(t) = \int_{r_w}^{\infty} \frac{\alpha q_t(r,t)}{h\lambda_t(r,t)k(r)} \frac{dr}{r} - \int_{r_w}^{r_{wf(t)}} f_w \frac{dp_c}{dS_w}\frac{\partial S_w}{\partial r}\, dr, \tag{52}$$

by assuming that the second term in the right-hand side of Eq. (52) is zero from r_{wf} to ∞, considering $f_w(S_{iw}) = 0$ and $\frac{\partial S_w}{\partial r} = 0$ for $r > r_{wf}(t)$, since the water in the region ahead of the water front foot is assumed immobile. r_{wf} can be defined as the position at which $(S_w - S_{iw}) < \delta$, where δ is a very small number. For the hypodispersion phenomenon, we can find a finite r_{wf} where $\delta \to 0$, i.e., at which $S_w = S_{iw}$. Using the [6] steady-state theory, which assumes that, $q_t(r, t) = q_t(r_w, t)$, for $r \le r_{wf}(t)$, Eq. (52) becomes

$$\Delta p_{wf}(t) = \frac{\alpha q_t(r_w,t)}{h}\int_{r_w}^{r_{wf}(t)} \frac{1}{\lambda_t\left(r,\Delta t_{prod}\right)k(r)}\frac{dr}{r} + \frac{\alpha}{h}\int_{r_{wf}(t)}^{\infty} \frac{q_t(r,t)}{\lambda_t(r,t)k(r)}\frac{dr}{r} - \int_{r_w}^{r_{wf}(t)} f_w \frac{dp_c}{dS_w}\frac{\partial S_w}{\partial r}dr, \tag{53}$$

where for any practical set of values of physical properties [28] indicate that this assumption is valid. Adding and subtracting the term $\frac{\alpha}{h}\int_{r_w}^{r_{wf}(t)} \frac{q_t(r_w,t)}{\widehat{\lambda}_o k(r)}\frac{dr}{r}$, where $\widehat{\lambda}_o = \frac{k_{ro}(S_{wi})}{\mu_o}$ is the endpoint oil mobility at $S_w = S_{wi}$, Eq. (53) can be rewritten as

$$\begin{aligned}\Delta p_{wf}(t) = {} & \frac{\alpha}{h}\int_{r_w}^{\infty} \frac{q_t(r,t)}{\widehat{\lambda}_o}(r,t)k(r)\,\frac{dr}{r} + \frac{\alpha q_t(r_w,t)}{h}\int_{r_w}^{r_{wf}(t)} \left(\frac{1}{\lambda_t(r,t)} - \frac{1}{\widehat{\lambda}_o}\right)\frac{dr}{k(r)r} - \int_{r_w}^{r_{wf}(t)} f_w \frac{dp_c}{dS_w}\frac{\partial S_w}{\partial r}dr \\ = {} & \Delta\widehat{p}_o(t) + \frac{\alpha q_t(r_w,t)}{h\widehat{\lambda}_o}\int_{r_w}^{r_{wf}(t)} \left(\frac{\widehat{\lambda}_o}{\lambda_t(r,t)} - 1\right)\frac{1}{k(r)}\frac{dr}{r} - \int_{r_w}^{r_{wf}(t)} f_w \frac{dp_c}{dS_w}\frac{\partial S_w}{\partial r}dr. \end{aligned} \tag{54}$$

$\Delta\widehat{p}_o(t)$ is the single-phase oil transient pressure drop, the known pressure drop solution that is obtained if we inject oil into an oil reservoir (injection period), whose well-known approximate solution can be approximated as

$$\Delta\widehat{p}_o(t) = p_{wf,o}(t) - p_i = \frac{\alpha q_t}{kh\widehat{\lambda}_o}\left[\frac{1}{2}\ln\left(\frac{\beta k\widehat{\lambda}_o t}{\phi\widehat{c}_{to}r_w^2}\right) + 0.4045 + s\right]. \tag{55}$$

Here, β is a unit conversion factor in which oil field unit is 0.0002637 and the single-phase total compressibility is

$$\widehat{c}_{to} = c_o(1 - S_{wi}) + c_w S_{wi} + c_r. \tag{56}$$

3. Validation

We have compared our pressure and saturation solution including capillary pressure effects with the commercial numerical simulator IMEX, using the properties shown in **Table 1**. **Figure 12** compares the saturation distribution obtained from our analytical solution with the one obtained with IMEX, while **Figure 13** shows the comparison of the wellbore pressure response from our analytical solution and IMEX during injection test. In order to be able to match saturation and pressure obtained from our solution with IMEX, we have to use a very refined grid (0.01 ft) around the wellbore in the zone invaded by water and then increase it exponentially to a very large external radius (10,000 times the wellbore radius) in order to reproduce an infinite acting reservoir. In addition, we have to start with very short time steps, 10^{-7} day. **Figure 14** presents the log-log diagnostic plots of injection. We can see that at early times of injection, there is a plateau (stabilization) in the wellbore pressure derivative plot, which by inspection reflects the original total mobility (endpoint oil mobility), while, at late times, we find that the derivative shows stabilized radial flow, which by inspection reflects the endpoint water mobility.

Property	Value	Unit	Property	Value	Unit
q_t	3000	RB/DAY	B_o	1.003	RB/STB
h	60	ft	B_w	1.002	RB/STB
r_w	0.35	ft	c_o	8×10^{-6}	1/psi
r_e	6800	ft	c_w	3.02×10^{-6}	1/psi
k	300	ft	c_r	5×10^{-6}	1/psi
s	0	mD	μ_o	3.0	cp
S_{iw}	0.10		μ_w	0.5	cp
S_{or}	0.25		λ	2	
p_i	2500	psi	p_t	0.5	psi
ϕ	0.22				

Table 1. Reservoir, rock, and fluid properties for simulation and analytical solution.

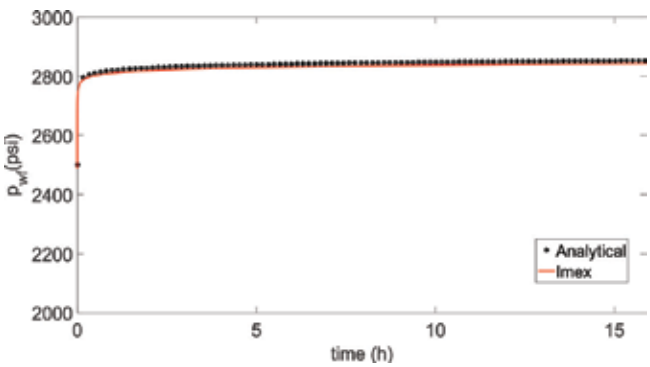

Figure 13. Comparison of the wellbore pressure response from analytical solution and IMEX during water injection test with capillary pressure (a).

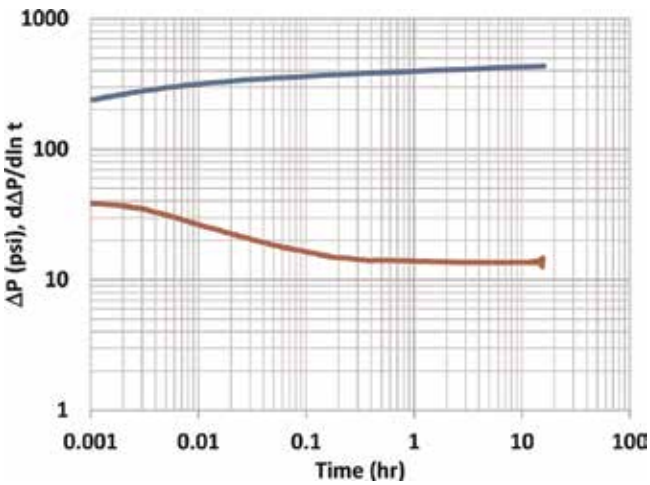

Figure 14. The log-log diagnostic plots for wellbore pressure data (blue curve) and its derivative with respect to time (red curve) during water injection (b).

http://dx.doi.org/10.5772/intechopen.76828

4. Conclusions

In this work, an accurate approximate analytical solution was constructed for wellbore pressure during water injection test in a reservoir containing oil and immobile water. Our solution was validated by comparing the bottom hole pressure calculated from the analytical model with the data obtained from a commercial numerical simulator. Our solution presented here for water injection together with the wellbore pressure and flow rate history for subsequent tests as shut-in and flowback can be used as forward model in a nonlinear regression in order to estimate relative permeabilities and capillary pressure curves in addition to the rock absolute permeability, the skin zone permeability, and the water endpoint relative permeability.

Acknowledgements

This research was conducted under the auspices of TUPREP, the Tulsa University Petroleum Reservoir Exploitation Projects, and it was prepared with financial support from the Coordination for the Improvement of Higher Education Personnel (CAPES) within the Brazilian Ministry of Education.

Conflict of interest

The authors declare that there is no conflict of interest.

Nomenclature

β Unit conversion factor (0.0002637)

Δp_o Single-phase oil pressure drop (psi)

λ_o Oil mobility (1/cp)

λ_t Total mobility (1/cp)

λ_w Water mobility (1/cp)

μ_o Oil viscosity (cp)

μ_w Water viscosity (cp)

a_w Water endpoint relative permeability

F_o Oil fractional flow

F_w Water fractional flow

f_w Water mobility ratio

k Absolute permeability (mD)

k_s Skin permeability (mD)

k_{ro} Oil relative permeability (mD)

k_{rw} Water relative permeability (mD)

p_c Capillary pressure (psi)

p_i Reservoir initial pressure (psi)

p_o Oil pressure (psi)

p_t Pressure threshold (psi)

p_w Water pressure (psi)

q_o Oil flow rate (RB/D)

q_t Total liquid rate (RB/D)

r_s Shock front position (ft)

r_{skin} Skin zone radius (ft)

r_w Wellbore radius (ft)

S_w Water saturation

S_{iw} Immobile water saturation

S_{or} Residual oil saturation

t Time (h)

C Constant given by $\frac{\theta q_t}{\pi h \phi}$

D Shock speed

h Reservoir thickness (ft)

r Radius (ft)

w Traveling wave coordinate

α Unit conversion factor (141.2)

ϵ Perturbation parameter

λ Pore size distribution index

ϕ Porosity

θ Unit conversion factor (5.6146/24)

c_o Oil compressibility (1/psi)

c_r Rock compressibility (1/psi)

c_w Water compressibility (1/psi)

c_{to} Single-phase total compressibility (1/psi)

Author details

Cíntia Gonçalves Machado* and Albert C. Reynolds

*Address all correspondence to: cintia-machado@utulsa.edu

McDougall School of Petroleum Engineering, University of Tulsa, Tulsa, OK, USA

References

[1] Rapoport LA, Leas WJ. Properties of linear waterfloods. Society of Petroleum Engineers. Journal of Petroleum Technology. 1953;**5**(5):139-148

[2] Peres AMM, Reynolds AC. Theory and analysis of injectivity tests on horizontal wells. SPE Journal. 2003;**8**(2):147-159

[3] Machado CG, Reynolds AC. Approximate semi-analytical solution for injection-falloff-production well test: An analytical tool for the in situ estimation of relative permeability curves. Transport in Porous Media. January 2018;**121**(1):207-231

[4] Machado CG, Firoozabad MM, Reynolds AC. Carbon dioxide effects on wellbore-pressure response during injection/falloff test. SPE Journal. January 2018. DOI: https://doi.org/10.2118/185824-PA

[5] Buckley SE, Leverett MC. Mechanism of fluid displacement in sands. SPE. Transactions of the AIME. December 1942;**146**(1):107-116

[6] Thompson LG, Reynolds AC. Well testing for radially heterogeneous reservoirs under single and multiphase flow conditions. SPEFE. 1997;**12**(1):57-64

[7] Thompson LG, Reynolds AC. Pressure transient analysis for gas condensate reservoirs. In Situ. 1997;**21**(2):101-144

[8] Brooks RH, Corey AT. Hydraulic properties of porous media. Hydrology Paper No 3. 1964

[9] Honarpour MM, Koederitz F, Herbert A. Relative Permeability of Petroleum Reservoirs. CRC press. 1986

[10] Li K. Generalized Capillary Pressure and Relative Permeability Model Inferred from Fractal Characterization of Porous Media. In: SPE Annual Technical Conference and Exhibition, 26-29 September, Houston, Texas. 2004. DOI: https://doi.org/10.2118/89874-MS

[11] Donaldson EC, Ewall N, Singh B. Characteristics of capillary pressure curves. Journal of Petroleum Science and Engineering. 1991;**6**(3):249-261

[12] Purcell WR. Capillary Pressures-their Measurement Using Mercury and the Calculation of Permeability. Journal of Petroleum Technology. 1949;**1**(2):39-48

[13] Li K, Horne RN. Numerical Simulation with Input Consistency between Capillary Pressure and Relative Permeability. In: SPE Reservoir Simulation Symposium, 3-5 February, Houston, Texas. 2003. DOI: https://doi.org/10.2118/79716-MS

[14] Bacri J, Rosen M, Salin D. Capillary hyperdiffusion as a test of wettability. EPL (Europhysics Letters). 1990;**11**(2):127

[15] Novy R, Toledo P, Davis H, Scriven L. Capillary dispersion in porous media at low wetting phase saturations. Chemical Engineering Science. 1989;**44**(9):1785-1797

[16] Yortsos Y, Fokas A. An analytical solution for linear waterflood including the effects of capillary pressure. Society of Petroleum Engineers Journal. 1983;**23**(01):115-124

[17] Barenblatt GI, Entov VM, Ryzhik VM. Theory of fluid flows through natural rocks. Boston: Kluwer Academic Publishers; 1989

[18] Bedrikovetsky P, Rodrigues JRP, Britto PRF. Analytical model for the waterflood honouring capillary pressure (with applications to laboratory studies). In: SPE Latin America/Caribbean Petroleum Engineering Conference, 23-26 April, Port-of-Spain, Trinidad. Society of Petroleum Engineers. January, 1996. DOI: https://doi.org/10.2118/36130-MS

[19] Deng L, King MJ. Capillary corrections to Buckley-Leverett flow. In: SPE Annual Technical Conference and Exhibition, 28-30 September, Houston, Texas, USA. Society of Petroleum Engineers. 2015. DOI: https://doi.org/10.2118/175150-MS

[20] Van Dyke M. Perturbation Methods in Fluid Mechanics/Annotated Edition. Vol. 75. Stanford, California: The Parabolic Press; 1975

[21] Holmes MH. Introduction to Perturbation Methods. Vol. 20. New York: Springer-Verlag Science & Business Media; 2012

[22] Rhee HK, Aris R, Amundson NR. First-Order Partial Differential Equations. Theory and Application of Single Equations. Vol. I. Mineola, New York: Dover Publications; 2001

[23] King M, Dunayevsky V. Why waterflood works: A linearized stability analysis. In: SPE Annual Technical Conference and Exhibition, 8-11 October, San Antonio, Texas. Society of Petroleum Engineers; 1989

[24] Hussain F, Cinar Y, Bedrikovetsky P. A semi-analytical model for two phase immiscible flow in porous media honouring capillary pressure. In: SPE Annual Technical Conference and Exhibition, 8-11 October, San Antonio, Texas. Transport in Porous Media. 2011;**92**:187-212

[25] Knobel R. An Introduction to the Mathematical Theory of Waves (Student Mathematical Library, V. 3). 1st ed. Rhode Island: American Mathematical Society; 1999

[26] Nayfeh AH. Introduction to Perturbation Techniques. New York: A Wiley-Interscience publication; 1981

[27] Settari A, Aziz K. A computer model for two-phase coning saturation. Society of Petroleum Engineers Journal. 1974;**14**(3):221-236

[28] Peres AMM, Reynolds AC. Theory and analysis of injectivity tests on horizontal wells. In: Proceedings of the SPE Annual Technical Conference and Exhibition. SPE 71582; 2001

Chapter 8

Perturbation Method for Solar/Infrared Radiative Transfer in a Scattering Medium with Vertical Inhomogeneity in Internal Optical Properties

Yi-Ning Shi, Feng Zhang, Jia-Ren Yan,
Qiu-Run Yu and Jiangnan Li

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.77147

Abstract

A new scheme based on perturbation method is presented to solve the problem of solar/infrared radiative transfer (SRT/IRT) in a scattering medium, in which the inherent optical properties (IOPs) are vertically inhomogeneous. The Eddington approximation for SRT and the two-stream approximation for IRT are used as the zeroth-order solution, and multiple-scattering effect of inhomogeneous IOPs is included in the first-order solution. Observations show that the stratocumulus clouds are vertically inhomogeneous, and the accuracy of SRT/IRT for stratocumulus clouds by different solutions is evaluated. In the spectral band of 0.25–0.69 μm, the relative error in absorption with inhomogeneous SRT solution is 1.4% at most, but with the homogeneous SRT solution, it can be up to 7.4%. In the spectral band of 5–8 μm, the maximum relative error of downward emissivity can reach −11% for the homogeneous IRT solution but only −2% for the inhomogeneous IRT solution.

Keywords: perturbation method, radiative transfer, vertical inhomogeneity

1. Introduction

Solving the radiative transfer equation (RTE) is a key issue in radiation scheme for climate model and remote sensing. In most numerical radiative transfer algorithms, the atmosphere is divided into many homogeneous layers. The inherent optical properties (IOPs) are then fixed within each layer and the variations of IOPs inside each layer are ignored, effectively regarding each layer as internally homogeneous. The standard solar/infrared radiative transfer (SRT/IRT)

solutions are based on this assumption of internal homogeneity [1–4], which cannot resolve within-layer vertical inhomogeneity.

It has been well established by observation that cumulus and stratocumulus clouds (hereinafter, collectively referred to as cumulus clouds) are inhomogeneous, both horizontally and vertically [5–9]. Inside a cumulus cloud, the liquid water content (LWC) and the cloud droplet size distribution vary with height, and so the IOPs of cloud droplets depend on vertical height.

How to deal with vertical internal inhomogeneity in SRT/IRT models is an interesting topic for researchers. Li developed a Monte Carlo cloud model that can be used to investigate photon transport in inhomogeneous clouds by considering an internal variation of the optical properties [10]. Their model showed that when overcast clouds become broken clouds, the difference in reflectance at large solar zenith angles between vertically inhomogeneous clouds and their plane-parallel counterparts can be as much as 10%.

However, the Monte Carlo method is very expensive in computing and not applicable to climate models or remote sensing [11]. The albedo of inhomogeneous mixed-phase clouds at visible wavelengths could be obtained by using a Monte Carlo method to compare such clouds with plane-parallel homogeneous clouds [12].

In principle, the vertical inhomogeneity problem of the SRT/IRT process can be solved by increasing the number of layers of the climate model. However, it is time-consuming to increase the vertical resolution of a climate model. Typically, there are only 30–100 layers in a climate model [13], which is not high enough to resolve the cloud vertical inhomogeneity. To completely address the problem of vertical inhomogeneity by using a limited number of layers in a climate model, the standard SRT method must be extended to deal with the vertical inhomogeneity inside each model layer. The primary purpose of this study is to introduce a new inhomogeneous SRT/IRT solution presented by Zhang and Shi. This solution follows a perturbation method: the zeroth-order solution is the standard Eddington approximation for SRT and two-stream approximation for IRT, with a first-order perturbation to account for the inhomogeneity effect. In Section 2, the basic theory of SRT/IRT is introduced, and the new inhomogeneous SRT/IRT solution is presented. In Section 3, the inhomogeneous SRT/IRT solution is applied to cloud as realistic examples to demonstrate the practicality of this new method. A summary is given in Section 4.

2. SRT/IRT solution for an inhomogeneous layer

2.1. SRT solution

The azimuthally averaged solar radiative transfer equation [1–4, 10–12] is

$$\mu \frac{dI_S(\tau,\mu)}{d\tau} = I_S(\tau,\mu) - \frac{\omega(\tau)}{2}\int_{-1}^{1} I_S(\tau,\mu)P(\tau,\mu,\mu')d\mu' - \frac{\omega(\tau)}{4\pi}F_0 P(\tau,\mu,-\mu_0)e^{-\frac{\tau}{\mu_0}} \qquad (1)$$

where μ is the cosine of the zenith angle ($\mu > 0$ and $\mu < 0$ refer to upward and downward radiation, respectively), $P(\tau,\mu,\mu')$ is the scattering phase function, τ is the optical depth

($\tau = 0$ and $\tau = \tau_0$ refer to the top and bottom of the medium, respectively), $\omega(\tau)$ is the single-scattering albedo, and F_0 is the incoming solar flux. For the Eddington approximation, $P(\tau, \mu, \mu') = 1 + 3g(\tau)\mu\mu'$ ($-1 < \mu < 1$) and $g(\tau)$ are the asymmetry factors. For the scattering atmosphere, the irradiance fluxes in the upward and downward directions can be written as

$$F_S^{\pm}(\tau) = 2\pi \int_0^{\pm 1} I_S(\tau, \mu)\mu d\mu \tag{2}$$

To simulate a realistic medium such as cloud or snow, we consider $\omega(\tau)$ and $g(\tau)$ to vary with τ, and we use exponential expressions here to simplify the process. The single-scattering albedo and asymmetry factor are written as

$$\omega(\tau) = \widehat{\omega} + \varepsilon_\omega \left(e^{-a_1\tau} - e^{-a_1\tau_0/2} \right) \tag{3a}$$

$$g(\tau) = \widehat{g} + \varepsilon_g \left(e^{-a_2\tau} - e^{-a_2\tau_0/2} \right) \tag{3b}$$

where τ_0 is the optical depth of the layer, $\widehat{\omega}$ is the single-scattering albedo at $\tau_0/2$, and $\widehat{g}$ is the asymmetry factor at the same place. Both ε_g and ε_ω are small parameters that are far less than $\widehat{g}$ and $\widehat{\omega}$, respectively, in a realistic medium.

According to the Eddington approximation, the radiative intensity $I_S(\tau, \mu)$ can be written as

$$I_S(\tau, \mu) = I_{S0}(\tau) + I_{S1}(\tau)\mu \tag{4}$$

Using Eqs. (1), (2), and (4), we obtain

$$\frac{dF_S^+(\tau)}{d\tau} = \gamma_1(\tau)F_S^+(\tau) - \gamma_2(\tau)F_S^-(\tau) - \gamma_3(\tau)\omega(\tau)F_0 e^{-\frac{\tau}{\mu_0}} \tag{5a}$$

$$\frac{dF_S^-(\tau)}{d\tau} = \gamma_2(\tau)F_S^+(\tau) - \gamma_1(\tau)F_S^-(\tau) + \left[1 - \gamma_3(\tau)\right]\omega(\tau)F_0 e^{-\frac{\tau}{\mu_0}} \tag{5b}$$

$$F_S^-(0) = 0, F_S^+(\tau_0) = R_{dif}F_S^-(\tau_0) + R_{dir}\mu_0 F_0 e^{-\frac{\tau_0}{\mu_0}} \tag{5c}$$

where $\gamma_1(\tau) = \frac{1}{4}\{7 - [4 + 3g(\tau)]\omega(\tau)\}$, $\gamma_2(\tau) = \frac{-1}{4}\{1 - [4 - 3g(\tau)]\omega(\tau)\}$, and $\gamma_3(\tau) = \frac{1}{4}\left[2 - 3g(\tau)\mu_0\right]$; τ_0 is the optical depth of the single layer; and R_{dif}(R_{dir}) is the diffuse (resp., direct) reflection from the layer below or the diffuse (direct) surface albedo. Substituting $\gamma_1(\tau)$, $\gamma_2(\tau)$, and $\gamma_3(\tau)$ into Eq. (3) and ignoring the small second-order parameters ${\varepsilon_\omega}^2$, ${\varepsilon_g}^2$, and $\varepsilon_\omega\varepsilon_g$, we get

$$\gamma_1(\tau) = \gamma_1^0 + \gamma_1^1\varepsilon_\omega \left(e^{-a_1\tau} - e^{-a_1\tau_0/2} \right) + \gamma_1^2\varepsilon_g \left(e^{-a_2\tau} - e^{-a_2\tau_0/2} \right) \tag{6a}$$

$$\gamma_2(\tau) = \gamma_2^0 + \gamma_2^1\varepsilon_\omega \left(e^{-a_1\tau} - e^{-a_1\tau_0/2} \right) + \gamma_2^2\varepsilon_g \left(e^{-a_2\tau} - e^{-a_2\tau_0/2} \right) \tag{6b}$$

$$\gamma_3(\tau) = \gamma_3^0 + \gamma_3^2\varepsilon_g \left(e^{-a_2\tau} - e^{-a_2\tau_0/2} \right) \tag{6c}$$

where $\gamma_1^0 = \frac{1}{4}[7-(4+3\widehat{g})\widehat{\omega}]$, $\gamma_2^0 = \frac{-1}{4}[1-(4-3\widehat{g})\widehat{\omega}]$, $\gamma_3^0 = \frac{1}{4}(2-3\widehat{g}\mu_0)$, $\gamma_1^1 = \frac{-1}{4}(4+3\widehat{g})$, $\gamma_1^2 = -\frac{3}{4}\widehat{\omega}$, $\gamma_2^1 = \frac{1}{4}(4-3\widehat{g})$, $\gamma_2^2 = -\frac{3}{4}\widehat{\omega}$, and $\gamma_3^2 = -\frac{3}{4}\mu_0$.

By perturbation theory [14], the corresponding flux can also be expanded by using the perturbation coefficients ε_ω and ε_g:

$$F_S^+ = F_{S0}^+ + \varepsilon_\omega F_{S1}^+ + \varepsilon_g F_{S2}^+ \tag{7a}$$

$$F_S^- = F_{S0}^- + \varepsilon_\omega F_{S1}^- + \varepsilon_g F_{S2}^- \tag{7b}$$

Substituting Eqs. (6) and (7) into Eq. (5) yields

$$\begin{aligned}\frac{dF_S^+}{d\tau} = &\left[\gamma_1^0 + \gamma_1^1\varepsilon_\omega\left(e^{-a_1\tau} - e^{-a_1\tau_0/2}\right) + \gamma_1^2\varepsilon_g\left(e^{-a_2\tau} - e^{-a_2\tau_0/2}\right)\right]\left(F_{S0}^+ + \varepsilon_\omega F_{S1}^+ + \varepsilon_g F_{S2}^+\right)\\ &-\left[\gamma_2^0 + \gamma_2^1\varepsilon_\omega\left(e^{-a_1\tau} - e^{-a_1\tau_0/2}\right) + \gamma_2^2\varepsilon_g\left(e^{-a_2\tau} - e^{-a_2\tau_0/2}\right)\right]\left(F_{S0}^- + \varepsilon_\omega F_{S1}^- + \varepsilon_g F_{S2}^-\right)\\ &-\left[\widehat{\omega}\gamma_3^0 + \gamma_3^0\varepsilon_\omega\left(e^{-a_1\tau} - e^{-a_1\tau_0/2}\right) + \widehat{\omega}\gamma_3^2\varepsilon_g\left(e^{-a_2\tau} - e^{-a_2\tau_0/2}\right)\right]F_0 e^{-\frac{\tau}{\mu_0}}\end{aligned} \tag{8a}$$

$$\begin{aligned}\frac{dF_S^-}{d\tau} = &\left[\gamma_2^0 + \gamma_2^1\varepsilon_\omega\left(e^{-a_1\tau} - e^{-a_1\tau_0/2}\right) + \gamma_2^2\varepsilon_g\left(e^{-a_2\tau} - e^{-a_2\tau_0/2}\right)\right]\left(F_{S0}^+ + \varepsilon_\omega F_{S1}^+ + \varepsilon_g F_{S2}^+\right)\\ &-\left[\gamma_1^0 + \gamma_1^1\varepsilon_\omega\left(e^{-a_1\tau} - e^{-a_1\tau_0/2}\right) + \gamma_1^2\varepsilon_g\left(e^{-a_2\tau} - e^{-a_2\tau_0/2}\right)\right]\left(F_{S0}^- + \varepsilon_\omega F_{S1}^- + \varepsilon_g F_{S2}^-\right)\\ &+\left[\widehat{\omega}\gamma_4^0 + \gamma_4^0\varepsilon_\omega\left(e^{-a_1\tau} - e^{-a_1\tau_0/2}\right) - \widehat{\omega}\gamma_3^2\varepsilon_g\left(e^{-a_2\tau} - e^{-a_2\tau_0/2}\right)\right]F_0 e^{-\frac{\tau}{\mu_0}}\end{aligned} \tag{8b}$$

where $\gamma_4^0 = 1 - \gamma_3^0$. And, Eq. (8) can be rewritten as separate equations for $F_{S0}^\pm$, $F_{S1}^\pm$, and $F_{S2}^\pm$. We obtain the following equations for the scattered flux $F_{S0}^\pm$:

$$\frac{dF_{S0}^+}{d\tau} = \gamma_1^0 F_{S0}^+ - \gamma_2^0 F_{S0}^- - \gamma_3^0\widehat{\omega}F_0 e^{-\frac{\tau}{\mu_0}} \tag{9a}$$

$$\frac{dF_{S0}^-}{d\tau} = \gamma_2^0 F_{S0}^+ - \gamma_1^0 F_{S0}^- + \gamma_4^0\widehat{\omega}F_0 e^{-\frac{\tau}{\mu_0}} \tag{9b}$$

$$F_{S0}^-(0) = 0, F_{S0}^+(\tau_0) = R_{dif}F_{S0}^-(\tau_0) + R_{dir}\mu_0 F_0 e^{-\frac{\tau_0}{\mu_0}} \tag{9c}$$

Eq. (9) is the standard SRT equation for a homogeneous layer [15] and has the following solution:

$$F_{S0}^+ = K_1 e^{k\tau} + \Gamma K_2 e^{-k\tau} + G_1 e^{-\frac{\tau}{\mu_0}} \tag{10a}$$

$$F_{S0}^- = \Gamma K_1 e^{k\tau} + K_2 e^{-k\tau} + G_2 e^{-\frac{\tau}{\mu_0}} \tag{10b}$$

where $K_1 = \frac{\left(\Gamma - R_{dif}\right)\Gamma G_2 e^{-k\tau_0} - \left(G_1 - R_{dir}G_2 - R_{dir}\mu_0 F_0\right)e^{-\frac{\tau_0}{\mu_0}}}{\left(1 - R_{dif}\Gamma\right)e^{k\tau_0} - \left(\Gamma - R_{dif}\right)\Gamma e^{-k\tau_0}}$, $K_2 = -\Gamma K_1 - G_2$, $G_1 = \left[\gamma_3^0\left(\frac{1}{\mu_0} - \gamma_1^0\right) - \gamma_2^0\gamma_4^0\right]\frac{{\mu_0}^2\widehat{\omega}F_0}{1-{\mu_0}^2k^2}$, $G_2 = -\left[\gamma_4^0\left(\frac{1}{\mu_0} + \gamma_1^0\right) + \gamma_2^0\gamma_3^0\right]\frac{{\mu_0}^2\widehat{\omega}F_0}{1-{\mu_0}^2k^2}$, $\Gamma = 1 - \frac{2k}{\gamma_1^0+\gamma_2^0+k}$, and $k^2 = \left(\gamma_1^0 + \gamma_2^0\right)\left(\gamma_1^0 - \gamma_2^0\right)$. And, the equations for the perturbation terms $F_{Si}^\pm$ (i = 1, 2) are

$$\frac{dF_{Si}^{+}}{d\tau}=\gamma_1^0F_{Si}^{+}-\gamma_2^0F_{Si}^{-}+\left(e^{-a_i\tau}-e^{-a_i\tau_0/2}\right)\left(\gamma_1^iF_{S0}^{+}-\gamma_2^iF_{S0}^{-}\right)-\gamma_3^{i-1}F_0\left(e^{-a_i\tau}-e^{-a_i\tau_0/2}\right)e^{-\frac{\tau}{\mu_0}} \tag{11a}$$

$$\frac{dF_{Si}^{-}}{d\tau}=\gamma_2^0F_{Si}^{+}-\gamma_1^0F_{Si}^{-}+\left(e^{-a_i\tau}-e^{-a_i\tau_0/2}\right)\left(\gamma_2^iF_{S0}^{+}-\gamma_1^iF_{S0}^{-}\right)-\gamma_4^{i-1}F_0\left(e^{-a_i\tau}-e^{-a_i\tau_0/2}\right)e^{-\frac{\tau}{\mu_0}} \tag{11b}$$

$$F_{Si}^{-}(0)=0,F_{Si}^{+}(\tau_0)=R_{dif}F_{Si}^{-}(\tau_0) \tag{11c}$$

where $\gamma_3^1=-\gamma_4^1=\widehat{\omega}\gamma_3^2$. Letting $M_i=F_{Si}^{+}+F_{Si}^{-}$ and $N_i=F_{Si}^{+}-F_{Si}^{-}$, Eq. (11a) and (11b) yields

$$\begin{aligned}\frac{dM_i}{d\tau}=&\left(\gamma_1^0+\gamma_2^0\right)N_i+\left(\psi_i^{+}+\psi_i^{-}\right)e^{-(k+a_i)\tau}+\left(\zeta_i^{+}+\zeta_i^{-}\right)e^{(k-a_i)\tau}+\left(\chi_i^{+}+\chi_i^{-}\right)e^{-\left(a_i+\frac{1}{\mu_0}\right)\tau}\\&-e^{-a_i\tau_0/2}\left[\left(\psi_i^{+}+\psi_i^{-}\right)e^{-k\tau}+\left(\zeta_i^{+}+\zeta_i^{-}\right)e^{k\tau}+\left(\chi_i^{+}+\chi_i^{-}\right)e^{-\frac{\tau}{\mu_0}}\right]\end{aligned} \tag{12a}$$

$$\begin{aligned}\frac{dN_i}{d\tau}=&\left(\gamma_1^0-\gamma_2^0\right)M_i+\left(\psi_i^{+}-\psi_i^{-}\right)e^{-(k+a_i)\tau}+\left(\zeta_i^{+}-\zeta_i^{-}\right)e^{(k-a_i)\tau}+\left(\chi_i^{+}-\chi_i^{-}\right)e^{-\left(a_i+\frac{1}{\mu_0}\right)\tau}\\&-e^{-a_i\tau_0/2}\left[\left(\psi_i^{+}-\psi_i^{-}\right)e^{-k\tau}+\left(\zeta_i^{+}-\zeta_i^{-}\right)e^{k\tau}+\left(\chi_i^{+}-\chi_i^{-}\right)e^{-\frac{\tau}{\mu_0}}\right]\end{aligned} \tag{12b}$$

where $\Psi_i^{+}=K_2\left(\gamma_1^i\Gamma-\gamma_2^i\right)$, $\Psi_i^{-}=K_2\left(\gamma_2^i\Gamma-\gamma_1^i\right)$, $\zeta_i^{+}=K_1\left(\gamma_1^i-\gamma_2^i\Gamma\right)$, $\zeta_i^{-}=K_1\left(\gamma_2^i-\gamma_1^i\Gamma\right)$, $\chi_i^{+}=\gamma_1^iG_1-\gamma_2^iG_2-\gamma_3^{i-1}F_0$, and $\chi_i^{-}=\gamma_2^iG_1-\gamma_1^iG_2+\gamma_4^{i-1}F_0$.

From Eq. (12), we obtain

$$\frac{d^2M_i}{d\tau^2}=k^2M_i+\eta_{1i}^{+}e^{-(k+a_i)\tau}+\eta_{2i}^{+}e^{(k-a_i)\tau}+\eta_{3i}^{+}e^{-\left(a_i+\frac{1}{\mu_0}\right)\tau}+\eta_{4i}^{+}e^{-k\tau}+\eta_{5i}^{+}e^{k\tau}+\eta_{6i}^{+}e^{-\frac{\tau}{\mu_0}} \tag{13a}$$

$$\frac{d^2N_i}{d\tau^2}=k^2N_i+\eta_{1i}^{-}e^{-(k+a_i)\tau}+\eta_{2i}^{-}e^{(k-a_i)\tau}+\eta_{3i}^{-}e^{-\left(a_i+\frac{1}{\mu_0}\right)\tau}+\eta_{4i}^{-}e^{-k\tau}+\eta_{5i}^{-}e^{k\tau}+\eta_{6i}^{-}e^{-\frac{\tau}{\mu_0}} \tag{13b}$$

where $\eta_{1i}^{\pm}=\left(\gamma_1^0\pm\gamma_2^0\right)\left(\psi_i^{+}\mp\psi_i^{-}\right)-(k+a_i)\left(\psi_i^{+}\pm\psi_i^{-}\right)$, $\eta_{2i}^{\pm}=(k-a_i)\left(\zeta_i^{+}\pm\zeta_i^{-}\right)+\left(\gamma_1^0\pm\gamma_2^0\right)\left(\zeta_i^{+}\mp\zeta_i^{-}\right)$, $\eta_{3i}^{\pm}=\left(\chi_i^{+}\mp\chi_i^{-}\right)\left(\gamma_1^0\pm\gamma_2^0\right)-\left(a_i+\frac{1}{\mu_0}\right)\left(\chi_i^{+}\mp\chi_i^{-}\right)$, $\eta_{4i}^{\pm}=-e^{-a_i\tau_0/2}\left[\left(\gamma_1^0\pm\gamma_2^0\right)\left(\psi_i^{+}\mp\psi_i^{-}\right)-k\left(\psi_i^{+}\pm\psi_i^{-}\right)\right]$, $\eta_{5i}^{\pm}=-e^{-a_i\tau_0/2}\left[k\left(\zeta_i^{+}\pm\zeta_i^{-}\right)+\left(\gamma_1^0\pm\gamma_2^0\right)\left(\zeta_i^{+}\mp\zeta_i^{-}\right)\right]$, and $\eta_{6i}^{\pm}=-e^{-a_i\tau_0/2}\left[\left(\chi_i^{+}\mp\chi_i^{-}\right)\left(\gamma_1^0\pm\gamma_2^0\right)-\frac{1}{\mu_0}\left(\chi_i^{+}\mp\chi_i^{-}\right)\right]$.

The solutions of Eq. (13) are

$$M_i=A_i^{+}e^{-k\tau}+B_i^{+}e^{k\tau}+P_i^{+}e^{-(k+a_i)\tau}+Q_i^{+}e^{(k-a_i)\tau}+R_i^{+}e^{-\left(a_i+\frac{1}{\mu_0}\right)\tau}-\frac{\eta_{4i}^{+}}{2k}e^{-k\tau}+\frac{\eta_{5i}^{+}}{2k}e^{k\tau}+\frac{\eta_{6i}^{+}\mu_0^2}{1-\mu_0^2k^2}e^{-\frac{\tau}{\mu_0}} \tag{14a}$$

$$N_i=A_i^{-}e^{-k\tau}+B_i^{-}e^{k\tau}+P_i^{-}e^{-(k+a_i)\tau}+Q_i^{-}e^{(k-a_i)\tau}+R_i^{-}e^{-\left(a_i+\frac{1}{\mu_0}\right)\tau}-\frac{\eta_{4i}^{-}}{2k}e^{-k\tau}+\frac{\eta_{5i}^{-}}{2k}e^{k\tau}+\frac{\eta_{6i}^{-}\mu_0^2}{1-\mu_0^2k^2}e^{-\frac{\tau}{\mu_0}} \tag{14b}$$

where $P_i^{\pm}=\frac{\eta_{1i}^{\pm}}{(k+a_i)^2-k^2}$, $Q_i^{\pm}=\frac{\eta_{2i}^{\pm}}{(k-a_i)^2-k^2}$, and $R_i^{\pm}=\frac{\eta_{3i}^{\pm}}{\left(a_i+\frac{1}{\mu_0}\right)^2-k^2}e^{-\frac{\tau}{\mu_0}}$. Finally, we can obtain F_{Si}^- and F_{Si}^+ as

$$F_{Si}^+=D_{1i}^+e^{-k\tau}+D_{2i}^+e^{k\tau}+\varphi_{1i}^+e^{-(k+a_i)\tau}+\varphi_{2i}^+e^{(k-a_i)\tau}+\varphi_{3i}^+e^{-\left(a_i+\frac{1}{\mu_0}\right)\tau}+\varphi_{4i}^+\tau e^{-k_1\tau}+\varphi_{5i}^+\tau e^{k\tau}+\varphi_{6i}^+e^{-\frac{\tau}{\mu_0}} \tag{15a}$$

$$F_{Si}^-=D_{1i}^-e^{-k\tau}+D_{2i}^-e^{k\tau}+\varphi_{1i}^-e^{-(k+a_i)\tau}+\varphi_{2i}^-e^{(k-a_i)\tau}+\varphi_{3i}^-e^{-\left(a_i+\frac{1}{\mu_0}\right)\tau}+\varphi_{4i}^-\tau e^{-k_1\tau}+\varphi_{5i}^-\tau e^{k\tau}+\varphi_{6i}^-e^{-\frac{\tau}{\mu_0}} \tag{15b}$$

where $D_{1i}^{\pm}=A_i^+\alpha^{\mp}\pm X_i$, $D_{2i}^{\pm}=B_i^+\alpha^{\pm}\pm Y$, $\alpha^{\pm}=\frac{1}{2}\left(1\pm\frac{k}{\gamma_1^0+\gamma_2^0}\right)$, $X_i=\frac{e^{-a_i\tau_0/2}\left(\psi_i^+\pm\psi_i^-\right)}{2\left(\gamma_1^0+\gamma_2^0\right)}-\frac{\eta_{4i}^+}{4k\left(\gamma_1^0+\gamma_2^0\right)}$,

$Y_i=\frac{e^{-a_i\tau_0/2}\left(\zeta_i^+\pm\zeta_i^-\right)}{2\left(\gamma_1^0+\gamma_2^0\right)}+\frac{\eta_{5i}^+}{4k\left(\gamma_1^0+\gamma_2^0\right)}$, $\phi_{1i}^{\pm}=\frac{1}{2}\left(P_i^+\pm P_i^-\right)$, $\phi_{2i}^{\pm}=\frac{1}{2}\left(Q_i^+\pm Q_i^-\right)$, $\phi_{3i}^{\pm}=\frac{1}{2}\left(R_i^+\pm R_i^-\right)$, $\phi_{4i}^{\pm}=-\frac{\eta_{4i}^+\pm\eta_{4i}^-}{4k}$, $\phi_{5i}^{\pm}=\frac{\eta_{5i}^+\pm\eta_{5i}^-}{4k}$, and $\phi_{6i}^{\pm}=\frac{\left(\eta_{6i}^+\pm\eta_{6i}^-\right)\mu_0^2}{1-\mu_0^2k^2}$. B_i and A_i are determined by the boundary conditions as

$$\begin{aligned}B_i^+ &= -\frac{\phi_{1i}^-+\phi_{2i}^-+\phi_{3i}^-+\phi_{6i}^-)(\alpha^--R_{dif}\alpha^+)e^{-k\tau_0}+\alpha^+\left[\left(\phi_{1i}^+-R_{dif}\phi_{1i}^-\right)e^{-(k+a_i)\tau_0}+\left(\phi_{2i}^+-R_{dif}\phi_{2i}^-\right)e^{(k-a_i)\tau_0}\right]}{\alpha^-\left(\alpha^--R_{dif}\alpha^+\right)e^{-k\tau_0}-\alpha^+\left(\alpha^+-R_{dif}\alpha^-\right)e^{k\tau_0}}\\ &+\frac{\alpha^+\left[\phi_{3i}^+-R_{dif}\phi_{3i}^-\right)e^{-\left(a_i+\frac{1}{\mu_0}\right)\tau_0}+\phi_{4i}^+-R_{dif}\phi_{4i}^-)\tau_0e^{-k\tau_0}+\left(\phi_{5i}^+-R_{dif}\phi_{5i}^-\right)\tau_0e^{k\tau_0}+\left(\phi_{6i}^+-R_{dif}\phi_{6i}^-\right)e^{-\frac{\tau_0}{\mu_0}}\right]}{\alpha^-\left(\alpha^--R_{dif}\alpha^+\right)e^{-k\tau_0}-\alpha^+\left(\alpha^+-R_{dif}\alpha^-\right)e^{k\tau_0}}\\ &+\frac{(X_i+Y_i)\left(\alpha^--R_{dif}\alpha^+\right)e^{-k\tau_0}+\left(X_i+R_{dif}X_i\right)\alpha^+e^{-k\tau_0}+\left(Y_i+R_{dif}Y_i\right)\alpha^+e^{k\tau_0}}{\alpha^-\left(\alpha^--R_{dif}\alpha^+\right)e^{-k\tau_0}-\alpha^+\left(\alpha^+-R_{dif}\alpha^-\right)e^{k\tau_0}}\end{aligned} \tag{16a}$$

$$A_i^+=\frac{1}{\alpha^+}\left(X_i+Y_i-B_i\alpha^--\phi_{1i}^--\phi_{2i}^--\phi_{3i}^--\phi_{6i}^-\right) \tag{16b}$$

All detailed calculation about solar radiation can be found at [16].

2.2. IRT solution

The azimuthally averaged infrared radiative transfer equation for intensity $I_1(\tau,\mu)$ is [1–4, 10–12]

$$\mu\frac{dI_1(\tau,\mu)}{d\tau}=I_I(\tau,\mu)-\frac{\omega(\tau)}{2}\int_{-1}^{1}I_1(\tau,\mu)P\left(\tau,\mu,\mu'\right)d\mu'-[1-\omega(\tau)]B(T) \tag{17}$$

where μ, τ, $P\left(\tau,\mu,\mu'\right)$, and $\omega(\tau)$ are same as in Eq. (1). $B(T)$ is the Planck function at temperature T, which represents the internal infrared emission of the medium.

The Planck function is approximated lineally as a function of optical depth [2] as

$$B[T(\tau)] = B_0 + \beta\tau \tag{18}$$

where $\beta = (B_1 - B_0)/\tau_0$ and τ_0 are the total optical depth of the medium. The Planck functions B_0 and B_1 are evaluated by using the temperature of the top ($\tau = 0$) and the bottom ($\tau = \tau_0$) of the medium.

According to the two-stream approximation, the intensities can be written as $I_I(\tau, \mu_1) = I_I^+(\tau)$ and $I_I(\tau, \mu_{-1}) = I_I^-(\tau)$, respectively, where $\mu_1 = -\mu_{-1} = 1/1.66$ is a diffuse factor that converts radiative intensity to flux [17]. $\int_{-1}^{1} I_I(\tau,\mu)P(\tau,\mu,\mu')d\mu'$ can be written as

$$\int_{-1}^{1} I_I(\tau,\mu)P(\tau,\mu,\mu')d\mu' = \left[1 + 3g(\tau)\mu\mu_1\right]I_I^+(\tau) + \left[1 + 3g(\tau)\mu\mu_{-1}\right]I_I^-(\tau) \tag{19}$$

where $g(\tau)$ is the asymmetry factor.

Using Eqs. (17) and (19), we can obtain

$$\frac{dI_I^+(\tau)}{d\tau} = \gamma_1(\tau)I_I^+(\tau) - \gamma_2(\tau)I_I^-(\tau) - \gamma_3(\tau)B(\tau) \tag{20a}$$

$$\frac{dI_I^+(\tau)}{d\tau} = \gamma_2(\tau)I_I^+(\tau) - \gamma_1(\tau)I_I^-(\tau) + \gamma_3(\tau)B(\tau) \tag{20b}$$

where $\gamma_1(\tau) = \frac{1-\omega(\tau)(1+g(\tau))/2}{\mu_1}$, $\gamma_2(\tau) = \frac{\omega(\tau)[1-g(\tau)]}{2\mu_1}$, and $\gamma_3(\tau) = \frac{1-\omega(\tau)}{\mu_1}$.

For IRT, we also use Eq. (3) to represent an inhomogeneous medium such as cloud or snow, in which $\omega(\tau)$ and $g(\tau)$ vary with τ. By substituting Eq. (3) into $\gamma_1(\tau)$, $\gamma_2(\tau)$, and $\gamma_3(\tau)$ and by ignoring the second order of the small parameters of ε_ω^2, ε_g^2, and $\varepsilon_\omega\varepsilon_g$, we can obtain

$$\gamma_1(\tau) = \gamma_1^0 + \gamma_1^1\varepsilon_\omega\left(e^{-a_1\tau} - e^{-a_1\tau_0/2}\right) + \gamma_1^2\varepsilon_g\left(e^{-a_2\tau} - e^{-a_2\tau_0/2}\right) \tag{21a}$$

$$\gamma_2(\tau) = \gamma_2^0 + \gamma_2^1\varepsilon_\omega\left(e^{-a_1\tau} - e^{-a_1\tau_0/2}\right) + \gamma_2^2\varepsilon_g\left(e^{-a_2\tau} - e^{-a_2\tau_0/2}\right) \tag{21b}$$

$$\gamma_3(\tau) = \gamma_3^0 + \gamma_3^1\varepsilon_\omega\left(e^{-a_1\tau} - e^{-a_1\tau_0/2}\right) \tag{21c}$$

In the above formula, γ_i^0, γ_i^1, and γ_i^2 (i = 1, 2, 3) are the known factors of $\widehat{\omega}$ and $\widehat{g}$. These known factors are introduced for simplifying original expressions, in which $\gamma_1^0 = \frac{1-\widehat{\omega}\left(1+\widehat{g}\right)/2}{\mu_1}$, $\gamma_2^0 = \frac{\widehat{\omega}\left(1-\widehat{g}\right)}{2\mu_1}$, $\gamma_3^0 = \frac{1-\widehat{\omega}}{\mu_1}$, $\gamma_1^1 = -\frac{1+\widehat{g}}{2\mu_1}$, $\gamma_2^1 = \frac{1-\widehat{g}}{2\mu_1}$, $\gamma_3^1 = -\frac{1}{\mu_1}$, $\gamma_1^2 = \gamma_2^2 = -\frac{\widehat{\omega}}{2\mu_1}$, and $\gamma_3^2 = 0$.

Same as in Eq. (7), the upward and downward intensity can be written as

$$I_I^+ = I_{I0}^+ + \varepsilon_\omega I_{I1}^+ + \varepsilon_g I_{I1}^+ \tag{22a}$$

$$I_I^- = I_{I0}^- + \varepsilon_\omega I_{I1}^- + \varepsilon_g I_{I1}^- \tag{22b}$$

By substituting Eqs. (21)–(22) into Eq. (20), we obtain

$$\begin{aligned}\frac{dI_{\mathrm{I}}^{+}}{d\tau} &= \left[\gamma_1^0+\gamma_1^1\varepsilon_\omega\left(e^{-a_1\tau}-e^{-a_1\tau_0/2}\right)+\gamma_1^2\varepsilon_g\left(e^{-a_2\tau}-e^{-a_2\tau_0/2}\right)\right]\left(I_{\mathrm{I}0}^{+}+\varepsilon_\omega I_{\mathrm{I}1}^{+}+\varepsilon_g I_{\mathrm{I}2}^{+}\right)\\ &\quad-\left[\gamma_2^0+\gamma_2^1\varepsilon_\omega\left(e^{-a_1\tau}-e^{-a_1\tau_0/2}\right)+\gamma_2^2\varepsilon_g\left(e^{-a_2\tau}-e^{-a_2\tau_0/2}\right)\right]\left(I_{\mathrm{I}0}^{-}+\varepsilon_\omega I_{\mathrm{I}1}^{-}+\varepsilon_g I_{\mathrm{I}2}^{-}\right)\\ &\quad-\left[\gamma_3^0+\gamma_3^1\varepsilon_\omega\left(e^{-a_1\tau}-e^{-a_1\tau_0/2}\right)\right]B(\tau)\end{aligned} \tag{23a}$$

$$\begin{aligned}\frac{dI_{\mathrm{I}}^{-}}{d_\tau} &= \left[\gamma_2^0+\gamma_2^1\varepsilon_\omega\left(e^{-a_1\tau}-e^{-a_1\tau_0/2}\right)+\gamma_2^2\varepsilon_g\left(e^{-a_2\tau}-e^{-a_2\tau_0/2}\right)\right]\left(I_{\mathrm{I}0}^{+}+\varepsilon_\omega I_{\mathrm{I}1}^{+}+\varepsilon_g I_{\mathrm{I}2}^{+}\right)\\ &\quad-\left[\gamma_1^0+\gamma_1^1\varepsilon_\omega\left(e^{-a_1\tau}-e^{-a_1\tau_0/2}\right)+\gamma_1^2\varepsilon_g\left(e^{-a_2\tau}-e^{-a_2\tau_0/2}\right)\right]\left(I_{\mathrm{I}0}^{-}+\varepsilon_\omega I_{\mathrm{I}1}^{-}+\varepsilon_g I_{\mathrm{I}2}^{-}\right)\\ &\quad+\left[\gamma_3^0+\gamma_3^1\varepsilon_\omega\left(e^{-a_1\tau}-e^{-a_1\tau_0/2}\right)\right]B(\tau)\end{aligned} \tag{23b}$$

By removing the second-order and higher-order perturbation terms, we can also separate Eq. (23) into three equations of $I_{\mathrm{I}i}^{\pm}$ (i = 0, 1, 2). The equations of $I_{\mathrm{I}0}^{\pm}$ can be written as

$$\frac{dI_{\mathrm{I}0}^{+}}{d\tau}=\gamma_1^0 dI_{\mathrm{I}0}^{+}-\gamma_2^0 dI_{\mathrm{I}0}^{-}-\gamma_3^0 B(\tau) \tag{24a}$$

$$\frac{dI_{\mathrm{I}0}^{-}}{d\tau}=\gamma_2^0 I_{\mathrm{I}0}^{+}-\gamma_1^0 I_{\mathrm{I}0}^{-}+\gamma_3^0 B(\tau) \tag{24b}$$

$$I_{\mathrm{I}0}^{-}(0)=0, I_{\mathrm{I}0}^{+}(\tau_0)=(1-\varepsilon_s)I_{\mathrm{I}0}^{-}(\tau_0)+\varepsilon_s B(T_s) \tag{24c}$$

where T_s and ε_s are surface temperature and surface emissivity, respectively. Eq. (24) is the standard homogeneous two-stream infrared radiative transfer equation [3, 15] with solutions

$$I_{\mathrm{I}0}^{+}=\alpha^{+}K_0e^{-k(\tau_0-\tau)}+\alpha^{-}H_0e^{-k\tau}+G_1\tau+G_2^{+} \tag{25a}$$

$$I_{\mathrm{I}0}^{-}=\alpha^{-}K_0e^{-k(\tau_0-\tau)}+\alpha^{+}H_0e^{-k\tau}+G_1\tau+G_2^{-} \tag{25b}$$

where $k^2=\left(\gamma_1^0+\gamma_2^0\right)\left(\gamma_1^0-\gamma_2^0\right)$, $\alpha^{\pm}=\frac{1}{2}\left(1\pm\frac{k}{\gamma_1^0+\gamma_2^0}\right)$, $G_1=\frac{\gamma_3^0}{\gamma_1^0-\gamma_2^0}\beta$, $G_2^{\pm}=\frac{\gamma_3^0}{\gamma_1^0-\gamma_2^0}B_0\pm\frac{\beta\gamma_3^0}{k^2}$, $H_0=\frac{\alpha^{-}e^{-k\tau_0}\left[\left(G_2^{+}-RG_2^{-}\right)+(1-R)G_1\tau_0-(1-R)B(T_s)\right]-(\alpha^{+}-R\alpha^{-})G_2^{-}}{\alpha^{+}(\alpha^{+}-R\alpha^{-})-\alpha^{-}\left(\alpha^{-}-R_{dif}\alpha^{+}\right)e^{-2k\tau_0}}$, $K_0=-\frac{\alpha^{+}H_0+G_2^{-}}{\alpha^{-}e^{-k\tau_0}}$, and $R=1-\varepsilon_s$.

The equations for $I_{\mathrm{I}i}^{\pm}$ (i = 1, 2) are

$$\frac{dI_{\mathrm{I}i}^{+}}{d\tau}=\gamma_1^0 I_{\mathrm{I}i}^{+}-\gamma_2^0 I_{\mathrm{I}i}^{-}+\left(e^{-a_i\tau}-e^{-a_i\tau_0/2}\right)\left[\gamma_1^i I_{\mathrm{I}0}^{+}-\gamma_2^i I_{\mathrm{I}0}^{-}-\gamma_3^i B(\tau)\right] \tag{26a}$$

$$\frac{dI_{\mathrm{I}i}^{-}}{d\tau}=\gamma_2^0 I_{\mathrm{I}i}^{+}-\gamma_1^0 I_{\mathrm{I}i}^{-}+\left(e^{-a_i\tau}-e^{-a_i\tau_0/2}\right)\left[\gamma_2^i I_{\mathrm{I}0}^{+}-\gamma_1^i I_{\mathrm{I}0}^{-}+\gamma_3^i B(\tau)\right] \tag{26b}$$

$$I_{\mathrm{I}i}^{-}(0)=0, I_{\mathrm{I}i}^{+}(\tau_0)=(1-\varepsilon_s)I_{\mathrm{I}i}^{-}(\tau_0) \tag{26c}$$

Let $M_i=I_{\mathrm{I}i}^{+}+I_{\mathrm{I}i}^{-}$ and $N_i=I_{\mathrm{I}i}^{+}-I_{\mathrm{I}i}^{-}$. Eq. (26a) and (26b) yields

$$\frac{dM_i}{d\tau} = \left(\gamma_1^0 + \gamma_2^0\right)N_i + \chi_{1i}^+ e^{-k\tau_0+(k-a_i)\tau} + \chi_{2i}^+ e^{-(k+a_i)\tau} + \chi_{3i}^+ e^{-k(\tau_0-\tau)} + \chi_{4i}^+ e^{-k\tau} + \chi_{5i}^+ + \chi_{6i}^+ e^{-a_i\tau} \quad (27a)$$

$$\frac{dN_i}{d\tau} = \left(\gamma_1^0 - \gamma_2^0\right)M_i + \chi_{1i}^- e^{-k\tau_0+(k-a_i)\tau} + \chi_{2i}^- e^{-(k+a_i)\tau} + \chi_{3i}^- e^{-k(\tau_0-\tau)} + \chi_{4i}^- e^{-k\tau} + \chi_{5i}^- + \chi_{6i}^- e^{-a_i\tau} + \chi_{7i}^-\tau + \chi_{8i}^-\tau e^{-a_i\tau} \quad (27b)$$

where $\chi_{1i}^{\pm} = K_0(\alpha^+ \mp \alpha^-)\left(\gamma_1^i \pm \gamma_2^i\right)$, $\chi_{2i}^{\pm} = \mp H_0(\alpha^+ \mp \alpha^-)\left(\gamma_1^i \pm \gamma_2^i\right)$, $\chi_{3i}^{\pm} = -K_0(\alpha^+ \mp \alpha^-)\left(\gamma_1^i \pm \gamma_2^i\right)e^{-a_i\tau_0/2}$, $\chi_{4i}^{\pm} = \pm H_0(\alpha^+ \mp \alpha^-)\left(\gamma_1^i \pm \gamma_2^i\right)e^{-a_i\tau_0/2}$, $\chi_{5i}^{+} = -\left(G_2^+ - G_2^-\right)\left(\gamma_1^i + \gamma_2^i\right)e^{-a_i\tau_0/2}$, $\chi_{5i}^{-} = -\left(G_2^+ + G_2^-\right)\left(\gamma_1^i - \gamma_2^i\right)e^{-a_i\tau_0/2} + 2B_0\gamma_3^i e^{-a_i\tau_0/2}$, $\chi_{6i}^{+} = \left(G_2^+ - G_2^-\right)\left(\gamma_1^i + \gamma_2^i\right)$, $\chi_{6i}^{-} = \left(G_2^+ + G_2^-\right)\left(\gamma_1^i - \gamma_2^i\right) - 2B_0\gamma_3^i$, $\chi_{7i}^{-} = \left[-2G_1\left(\gamma_1^i - \gamma_2^i\right) + 2\beta\gamma_3^i\right]e^{-a_i\tau_0/2}$, and $\chi_{8i}^{-} = 2G_1\left(\gamma_1^i - \gamma_2^i\right) - 2\beta\gamma_3^i$.

From Eq. (27), we can obtain

$$\frac{d^2M_i}{d\tau^2} = k^2M_i + \phi_{1i}^+ e^{-k\tau_0+(k-a_i)\tau} + \phi_{2i}^+ e^{-(k+a_i)\tau} + \phi_{3i}^+ e^{-k(\tau_0-\tau)} + \phi_{4i}^+ e^{-k\tau} + \phi_{5i}^+ + \phi_{6i}^+ e^{-a_i\tau} + \phi_{7i}^+\tau + \phi_{8i}^+\tau e^{-a_i\tau} \quad (28a)$$

$$\frac{d^2N_i}{d\tau^2} = k^2N_i + \phi_{1i}^- e^{-k\tau_0+(k-a_i)\tau} + \phi_{2i}^- e^{-(k+a_i)\tau} + \phi_{3i}^- e^{-k(\tau_0-\tau)} + \phi_{4i}^- e^{-k\tau} + \phi_{5i}^- + \phi_{6i}^- e^{-a_i\tau} + \phi_{8i}^-\tau e^{-a_i\tau} \quad (28b)$$

where $\phi_{1i}^{\pm} = \left(\gamma_1^0 \pm \gamma_2^0\right)\chi_{1i}^{\mp} + (k - a_i)\chi_{1i}^{\pm}$, $\phi_{2i}^{\pm} = \left(\gamma_1^0 \pm \gamma_2^0\right)\chi_{2i}^{\mp} - (k + a_i)\chi_{2i}^{\pm}$, $\phi_{3i}^{\pm} = \left(\gamma_1^0 \pm \gamma_2^0\right)\chi_{3i}^{\mp} + k\chi_{3i}^{\pm}$, $\phi_{4i}^{\pm} = \left(\gamma_1^0 \pm \gamma_2^0\right)\chi_{4i}^{\mp} - k\chi_{4i}^{\pm}$, $\phi_{5i}^{+} = \left(\gamma_1^0 + \gamma_2^0\right)\chi_{5i}^{-}$, $\phi_{5i}^{-} = \left(\gamma_1^0 - \gamma_2^0\right)\chi_{5i}^{+} + \chi_{7i}^{-}$, $\phi_{6i}^{+} = \left(\gamma_1^0 + \gamma_2^0\right)\chi_{6i}^{-} - a_i\chi_{6i}^{+}$, $\phi_{6i}^{-} = \left(\gamma_1^0 - \gamma_2^0\right)\chi_{6i}^{+} - a_i\chi_{6i}^{-} + \chi_{8i}^{-}$, $\phi_{7i}^{+} = \left(\gamma_1^0 + \gamma_2^0\right)\chi_{7i}^{-}$, $\phi_{8i}^{+} = \left(\gamma_1^0 + \gamma_2^0\right)\chi_{8i}^{-}$, and $\phi_{8i}^{-} = -a_i\chi_{8i}^{-}$. Thus, the solutions are

$$M_i = K_{1i}e^{-k(\tau_0-\tau)} + H_{1i}e^{-k\tau} + P_{1i}^+ e^{-k\tau_0+(k-a_i)\tau} + P_{2i}^+ e^{-(k+a_i)\tau} + P_{3i}^+\tau e^{-k(\tau_0-\tau)} + P_{4i}^+\tau e^{-k\tau} + P_{5i}^+ + P_{6i}^+ e^{-a_i\tau} + P_{7i}^+\tau + P_{8i}^+\tau e^{-a_i\tau} \quad (29a)$$

$$N_i = K_{2i}e^{-k(\tau_0-\tau)} + H_{2i}e^{-k\tau} + P_{1i}^- e^{-k\tau_0+(k-a_i)\tau} + P_{2i}^- e^{-(k+a_i)\tau} + P_{3i}^-\tau e^{-k(\tau_0-\tau)} + P_{4i}^-\tau e^{-k\tau} + P_{5i}^- + P_{6i}^- e^{-a_i\tau} + P_{8i}^-\tau e^{-a_i\tau} \quad (29b)$$

where $P_{1i}^{\pm} = \frac{\phi_{1i}^{\pm}}{(k-a_i)^2-k^2}$, $P_{2i}^{\pm} = \frac{\phi_{2i}^{\pm}}{(k+a_i)^2-k^2}$, $P_{3i}^{\pm} = \frac{\phi_{3i}^{\pm}}{2k}$, $P_{4i}^{\pm} = -\frac{\phi_{4i}^{\pm}}{2k}$, $P_{5i}^{\pm} = -\frac{\phi_{5i}^{\pm}}{k^2}$, $P_{6i}^{\pm} = \frac{\phi_{6i}^{\pm}\left(a_i^2-k^2\right)+2a_i\phi_{8i}^{\pm}}{\left(a_i^2-k^2\right)^2}$, $P_{7i}^{+} = -\frac{\phi_{7i}^{+}}{k^2}$, and $P_{8i}^{+} = \frac{\phi_{8i}^{+}}{a_i^2-k^2}$.

The expressions of $I_{1i}^{\pm}$ are

$$I_{1i}^+ = D_{1i}^+ e^{-k(\tau_0-\tau)} + D_{2i}^+ e^{-k\tau} + \sigma_{1i}^+ e^{-k\tau_0+(k-a_i)\tau} + \sigma_{2i}^+ e^{-(k+a_i)\tau} + \sigma_{3i}^+\tau e^{-k(\tau_0-\tau)} + \sigma_{4i}^+\tau e^{-k\tau} + \sigma_{5i}^+ + \sigma_{6i}^+ e^{-a_i\tau} + \sigma_{7i}\tau + \sigma_{8i}^+\tau e^{-a_i\tau} \quad (30a)$$

$$I_{1i}^- = D_{1i}^- e^{-k(\tau_0-\tau)} + D_{2i}^- e^{-k\tau} + \sigma_{1i}^- e^{-k\tau_0+(k-a_i)\tau} + \sigma_{2i}^- e^{-(k+a_i)\tau} + \sigma_{3i}^-\tau e^{-k(\tau_0-\tau)} + \sigma_{4i}^-\tau e^{-k\tau} + \sigma_{5i}^- + \sigma_{6i}^- e^{-a_i\tau} + \sigma_{7i}\tau + \sigma_{8i}^-\tau e^{-a_i\tau} \quad (30b)$$

where $D_{1i}^{\pm} = K_{1i}\alpha^{\pm} \pm X_i$, $D_{2i}^{\pm} = H_{1i}\alpha^{\mp} \pm Y_i$, $X_i = \frac{P_{3i}^{+} - \chi_{3i}^{+}}{2\left(\gamma_1^0 + \gamma_2^0\right)}$, $Y_i = \frac{P_{4i}^{+} - \chi_{4i}^{+}}{2\left(\gamma_1^0 + \gamma_2^0\right)}$, $\sigma_{ji}^{\pm} = \frac{1}{2}\left(P_{ji}^{+} \pm P_{ji}^{-}\right)$. (j = 1, 2, 3, 4, 5, 6, 8), $\sigma_{7i} = \frac{1}{2}P_{7i}^{+}$, and K_{1i} and H_{1i} are determined by boundary conditions. By substituting Eq. (30) into the boundary conditions of Eq. (26c), we can obtain

$$H_{1i} = \frac{(\alpha^{+} - R\alpha^{-})\left(X_i e^{-k\tau_0} + Y_i - \sigma_{1i}^{-}e^{-k\tau_0} - \sigma_{2i}^{-} - \sigma_{5i}^{-} - \sigma_{6i}^{-}\right) + \alpha^{-}e^{-k\tau_0}\left[(R+1)X_i + (R+1)Y_i e^{-k\tau_0}\right]}{\alpha^{+}\left(\alpha^{+} - R_{dif}\alpha^{-}\right) - \alpha^{-}\left(\alpha^{-} - R_{dif}\alpha^{-}\right)e^{-2k\tau_0}}$$
$$- \frac{\alpha^{-}e^{-k\tau_0}\left[\left(R\sigma_{1i}^{-} - \sigma_{1i}^{+}\right)e^{-a_i\tau_0} + \left(R\sigma_{2i}^{-} - \sigma_{2i}^{+}\right)e^{-(k+a_i)\tau_0} + \left(R\sigma_{3i}^{-} - \sigma_{3i}^{+}\right)\tau_0 + \left(R\sigma_{4i}^{-} - \sigma_{4i}^{+}\right)\tau_0 e^{-k\tau_0}\right]}{\alpha^{+}\left(\alpha^{+} - R_{dif}\alpha^{-}\right) - \alpha^{-}\left(\alpha^{-} - R_{dif}\alpha^{-}\right)e^{-2k\tau_0}}$$
$$- \frac{\alpha^{-}e^{-k\tau_0}\left[\left(R\sigma_{5i}^{-} - \sigma_{5i}^{+}\right) + \left(R\sigma_{6i}^{-} - \sigma_{6i}^{+}\right)e^{-a_i\tau_0} + (R-1)\sigma_{7i}\tau_0 + \left(R\sigma_{8i}^{-} - \sigma_{8i}^{+}\right)\tau_0 e^{-a_i\tau_0}\right]}{\alpha^{+}\left(\alpha^{+} - R_{dif}\alpha^{-}\right) - \alpha^{-}\left(\alpha^{-} - R_{dif}\alpha^{-}\right)e^{-2k\tau_0}} \tag{31a}$$

$$K_{1i} = \frac{1}{\alpha^{-}e^{-k\tau_0}}\left(X_i e^{-k\tau_0} + Y_i - \alpha^{+}H_{1i} - \sigma_{1i}^{-}e^{-k\tau_0} - \sigma_{2i}^{-} - \sigma_{5i}^{-} - \sigma_{6i}^{-}\right) \tag{31b}$$

Finally, the upward and downward fluxes are obtained by

$$F_{\mathrm{I}}^{+}(0) = \pi I_{\mathrm{I}}^{+}(0) \tag{32a}$$

$$F_{\mathrm{I}}^{-}(\tau_0) = \pi I_{\mathrm{I}}^{-}(\tau_0) \tag{32b}$$

All detailed calculation about solar radiation can be found at [18].

3. Results and discussion

We apply the two schemes to idealized medium to investigate its accuracy, and the result has been shown on [16] and [18].

For true cloud medium, because ice clouds' optical properties strongly depend on the complex particle habits [19–21]. Therefore, we limit our discussion here to water cloud only. According to the observation, the internal LWC (g m^{-3}) and droplet radius of the cloud tend to increase with height [22]. To take this feature into account, LWC and droplet cross-sectional area (DCA; cm^{-2}, m^{-3}) should increase linearly from the cloud base to the position near the top of the cloud:

$$\mathrm{LWC} = 0.22 + 0.00008z \tag{33a}$$

$$\mathrm{DCA} = 100 + z \tag{33b}$$

where $0 < z < z_0$. The terms z and z_0 denote the height from the cloud base and the height of the cloud top, respectively. From Eq. (33a) to (33b), the cloud effective radius (r_e; μm) and liquid water path (LWP; g m^{-2}) can be obtained:

$$r_e(z) = \frac{3}{4\rho}\frac{\mathrm{LWC}}{\mathrm{DCA}}10^{10} \tag{34a}$$

$$\mathrm{LWP} = \int_0^{z_0} \mathrm{LWC}dz \tag{34b}$$

where ρ (g m^{-3}) is the liquid water density. In this case, LWC varies from 0.22 to 0.30 g m^{-3}, and r_e varies from 2.06 to 16.50 μm, in which both ranges are consistent with observation [23]. According to [24], we choose LWP = 260 (g m^{-2}) to represent low cloud. In the benchmark calculations, z_0 is divided into 100 internal homogeneous sub-layers, although other numbers can be chosen (e.g., 200). In principle, more internal sub-layers should result in more accurate results. We use 100 internal sub-layers throughout this study because having any more makes little difference to the calculated results. Using 100 sub-layers are sufficiently accurate to resolve the vertical internal inhomogeneity of the medium. We use the optical properties of a

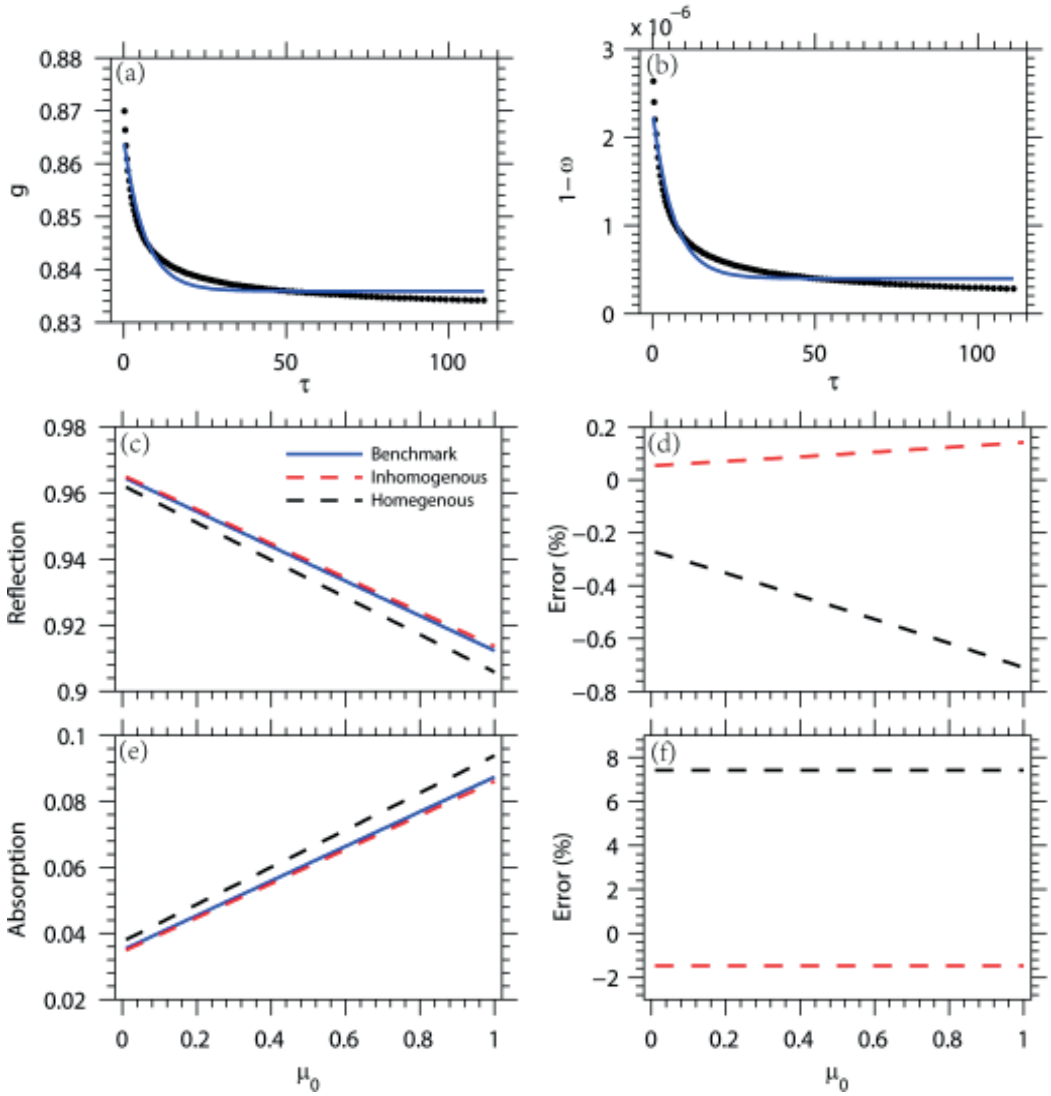

Figure 1. For the band of 0.25–0.69 μm, (a-b) show cloud asymmetry factor/single-scattering albedo versus cloud optical depth (a for asymmetry factor; b for single-scattering albedo), (c-d) show the reflectance/absorptance versus solar zenith angle (c for reflectance; d for absorptance) and (e-f) show the relative errors of the homogeneous and inhomogeneous solutions (e for reflectance error, f for absorptance error).

water cloud in the solar spectral band of 0.25–0.69 μm and at 0.94 μm and in the infrared spectral band of 5–8 μm and 11 μm.

In **Figure 1a** and **b**, the benchmark values of the inhomogeneous IOPs and the parameterized results for the spectral band of 0.25–0.69 μm are shown. The parameterized inhomogeneous IOPs are

$$1-\omega(\tau)=3.979\times10^{-7}-1.897\times10^{-6}\left(e^{-0.1539\tau}-e^{-0.1539\tau_0/2}\right) \tag{35a}$$

$$g(\tau)=0.8359+0.0289\left(e^{-0.1539\tau}-e^{-0.1539\tau_0/2}\right) \tag{35b}$$

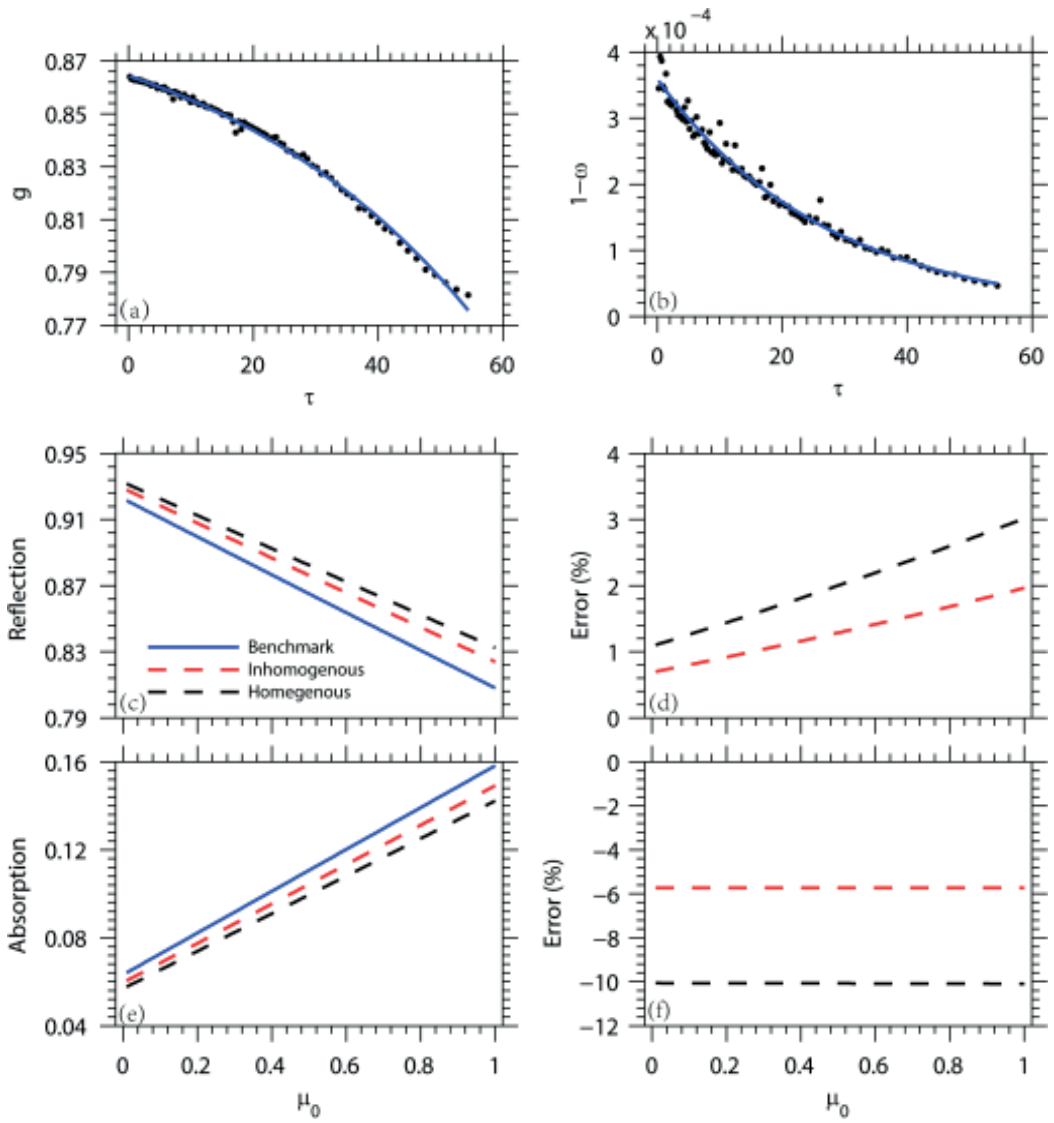

Figure 2. Same as **Figure 1** but for the wavelength 0.94 μm.

where $\tau_0 = 110.84$. The corresponding results for reflection and absorption are shown in **Figure 1c–f**. For reflection, the relative error with the homogeneous solution increases from 0.25 to 0.71% as μ_0 increases from 0.01 to 1, whereas the relative error with the inhomogeneous solution increases from 0.05 to 0.14%. For absorption, the relative error is not sensitive to μ_0; it is around 7.4% with the homogeneous solution but around only 1.4% with the inhomogeneous solution.

In **Figure 2a** and **b**, the benchmark values of the inhomogeneous IOPs and the parameterized results for the wavelength 0.94 μm are shown. The parameterized inhomogeneous IOPs are

$$1 - \omega(\tau) = 1.936 \times 10^{-4} - 5.263 \times 10^{-4}\left(e^{-0.0357\tau} - e^{-0.0357\tau_0/2}\right) \tag{36a}$$

$$g(\tau) = 0.8321 - 0.0403\left(e^{0.0218\tau} - e^{0.0218\tau_0/2}\right) \tag{36b}$$

where $\tau_0 = 54.46$. **Figure 2c–f** shows the corresponding results for reflection and absorption. For reflection, the relative error with the homogeneous solution increases from 1.1 to 3.0% as μ_0 increases from 0.01 to 1, whereas the relative error with the inhomogeneous solution increases from 0.7 to 2.0%. For absorption, the relative error is not sensitive to μ_0; it is around 10% with the homogeneous solution but around only 5.7% with the inhomogeneous solution.

The benchmark values of IOPs and parameterized results for the band of 5–8 μm are shown in **Figure 3a** and **b**. Here, we assume

$$\omega(\tau) = 0.6757 - 0.3697\left(e^{-0.0142\tau} - e^{-0.0142\tau_0/2}\right) \tag{37a}$$

$$g(\tau) = 0.8644 + 0.1023\left(e^{-0.0155\tau} - e^{0.0155\tau_0/2}\right) \tag{37b}$$

where $\tau_0 = 55.85$. For upward emissivity (**Figure 3c** and **d**), the relative errors of both solutions are not sensitive to $F_I^+(\tau_0)$; the errors are around -3% for homogeneous solution and around 1% for inhomogeneous solution. For downward emissivity (**Figure 3e** and **f**), the relative error of homogeneous solution is 4% when $F_I^+(\tau_0) = 0$, while the error of inhomogeneous solution is only 1%. With $F_I^+(\tau_0)$ increasing from 0 to $5\pi B(T)$, the error of homogeneous solution decreases to 0 firstly but then negatively increases to around -10%. The error of inhomogeneous solution shows a similar decreasing-increasing pattern, but the negative increase only reaches about -2%.

The benchmark values of IOPs and parameterized results for the band of 11 μm are shown in **Figure 4a** and **b**. In this case, we assume

$$\omega(\tau) = 0.4623 - 0.2155\left(e^{-0.1018\tau} - e^{-0.1018\tau_0/2}\right) \tag{38a}$$

$$g(\tau) = 0.9118 - 0.0083\left(e^{0.1087\tau} - e^{0.1087\tau_0/2}\right) \tag{38b}$$

where $\tau_0 = 28.23$. For upward emissivity (**Figure 4c** and **d**), the relative error of homogeneous solution is -1.2%, while the error of inhomogeneous solution is less than 0.5%. For downward emissivity (**Figure 4e** and **f**), with $F_{\mathrm{I}}^{+}(\tau_0)$ increasing from 0 to $5\pi B(T)$, the error of homogeneous (inhomogeneous) solution varies from 3 to -11% (from 0 to -1%).

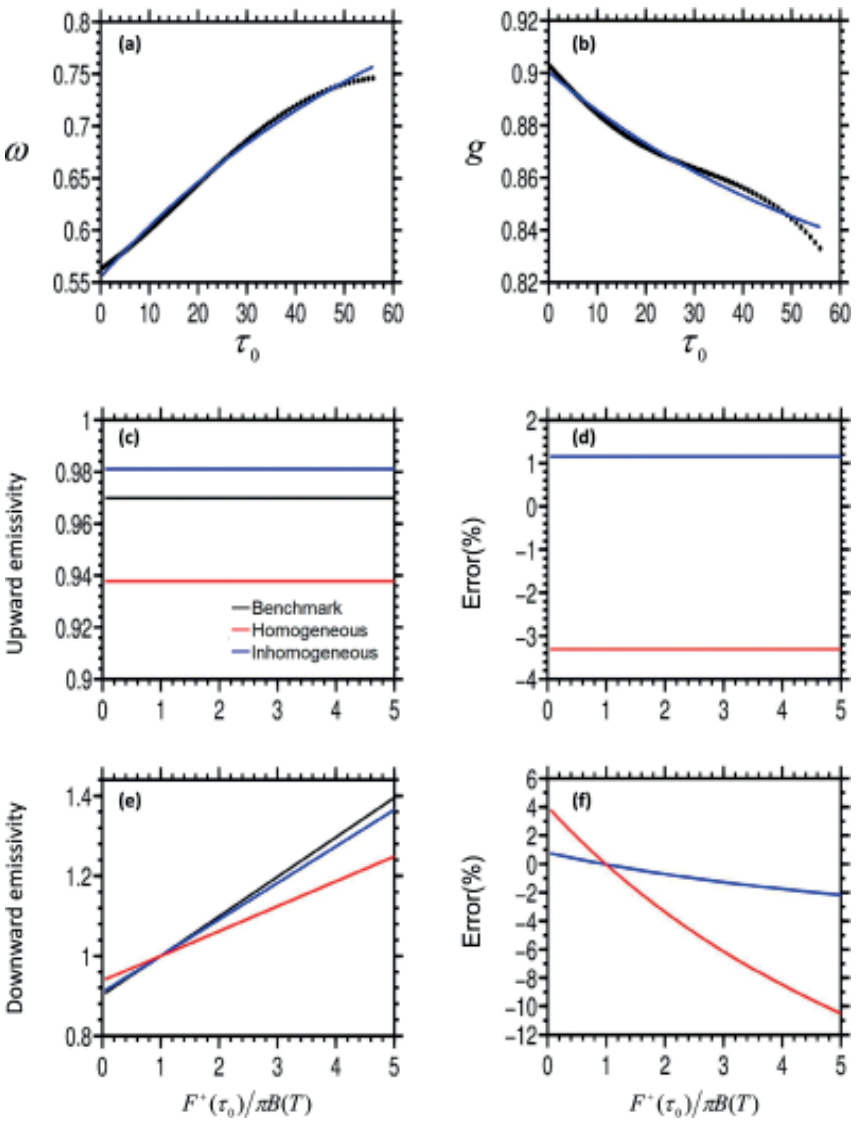

Figure 3. For the band of 5-8 μm, (a-b) show the cloud single-scattering albedo and asymmetry factor versus cloud optical depth, black dots represent the exact values and the blue lines is the fitting results (a for single-scattering albedo; b for asymmetry factor); (c-d) show the upward/downward emissivity versus the ratio of the radiation incident from the bottom to the internal infrared emission of the medium (c for upward emissivity; d for downward emissivity) and (e-f) show the relative errors of the homogeneous and inhomogeneous solutions (e for upward emissivity; f for downward emissivity).

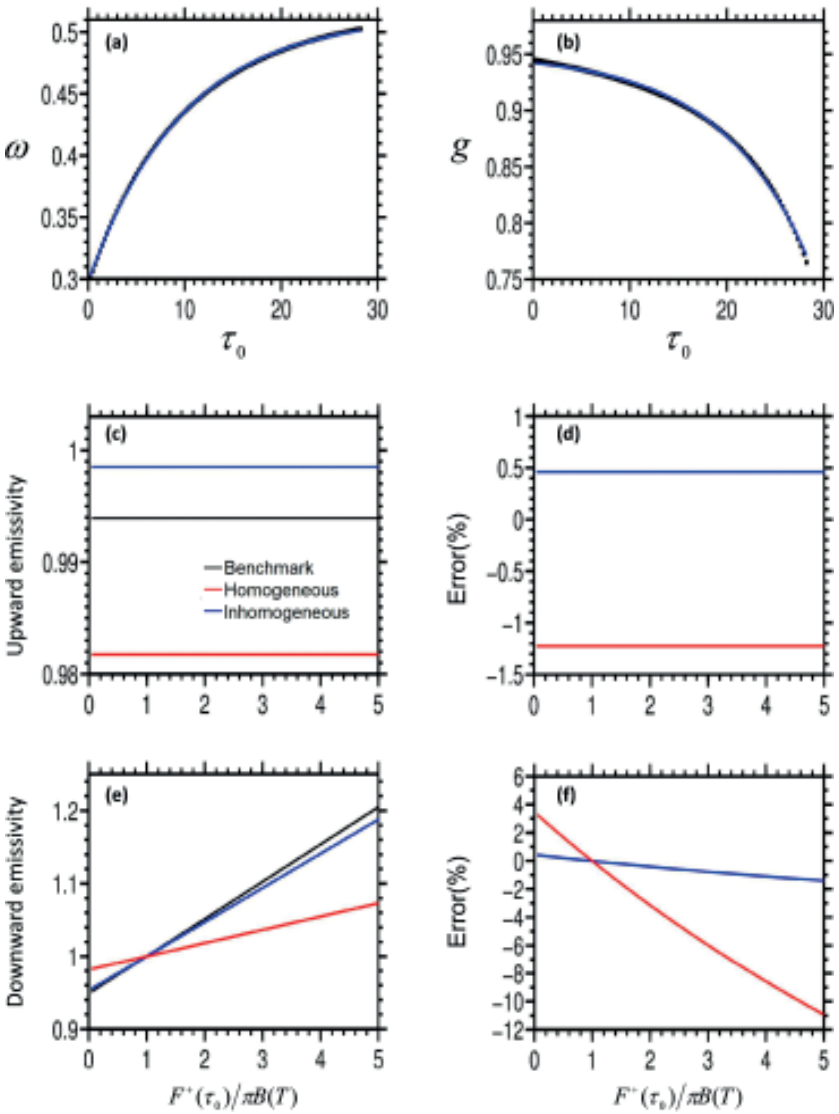

Figure 4. Same as **Figure 3** but for the wavelength of 11 μm.

4. Summary and conclusions

In the above, we have considered the vertically inhomogeneous structures of only cloud and snow, whereas all physical quantities in the atmosphere are vertically inhomogeneous (e.g., the concentrations of all types of gases and aerosols). In current climate models, the vertical layer resolution is far from that required to resolve such vertical inhomogeneity. In this study, we have proposed a new inhomogeneous SRT/IRT solution to address the vertical inhomogeneity

by introducing an internal variation of IOPs inside each model layer. This scheme is based on standard perturbation theory and allows us to use the standard solar Eddington solution and standard infrared two-stream solution for homogeneous layers to identify a zeroth-order equation and a first-order equation that includes the inhomogeneous effect. The new SRT/IRT solution can accurately express the inhomogeneous effect in each model layer, and it reduces to the standard solution when the medium is homogeneous.

The new inhomogeneous SRT/IRT solution is a good way to resolve cloud vertical inhomogeneity. In the spectral band of 0.25–0.69 μm, the relative error in the inhomogeneous SRT solution is no more than 1.4%, whereas the error with the homogeneous SRT solution can be up to 7.4%. At the specific wavelength of 0.94 μm, the relative error with the inhomogeneous solution is not more than 5.7% but can be up to 10% with the homogeneous SRT solution. In the band of 5–8 μm, the homogeneous IRT solution is not sensitive to $F_{\mathrm{I}}^{+}(\tau_0)$, and its relative error may reach -3.2% for upward emissivity, whereas the error of inhomogeneous IRT solution is only 1%. With $F_{\mathrm{I}}^{+}(\tau_0)$ increasing from 0 to $5\pi B(T)$, the error of downward emissivity for homogeneous solution varies from 4 to -10%, while the error ranges from 1 to -2% for inhomogeneous IRT solution. In the band of 11 μm, the relative error of homogeneous IRT solution is around -1.2% for upward emissivity, and the error of inhomogeneous IRT solution is only less than 0.5%. For downward emissivity, the maximum error of homogeneous IRT solution can be up to -11%, and the maximum error of inhomogeneous IRT solution is only around -1% when $F_{\mathrm{I}}^{+}(\tau_0) = 5\pi B(T)$.

In specific spectral bands or at particular wavelengths, the vertical variations in IOPs can typically be fitted easily into Eq. (3) to obtain the required parameters. A simple fitting program can be easily incorporated into a climate model to produce the inhomogeneous IOPs of stratocumulus clouds. If no such cloud inhomogeneity information is available in the current climate models, the vertical variation rates of cloud LWC and DCA can be derived empirically from observations, which show that the vertical variation rates of LWC and DCA in stratocumulus clouds are not very different [5, 7, 8].

In this study, we presented only a single-layer inhomogeneous SRT/IRT solution. To implement the new solution in a climate model, the adding process for layer-to-layer connections has to be solved. Under the homogeneous condition, the single-layer result in reflection and transmission is the same for an upward path and a downward path, but this is not true for an inhomogeneous layer. Therefore, the adding process has to be modified. We will present an algorithm for this multilayer adding process in our next study, in which the climatic impact of inhomogeneous clouds and inhomogeneous snows will be explored. The code base for the inhomogeneous SRT/IRT solution is available from the authors upon request.

Acknowledgements

The work is supported by National Natural Science Foundation of China (41675003) and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

Author details

Yi-Ning Shi[1], Feng Zhang[1]*, Jia-Ren Yan[1], Qiu-Run Yu[1] and Jiangnan Li[2]

*Address all correspondence to: fengzhang@nuist.edu.cn

1 Key Laboratory of Meteorological Disaster, Ministry of Education, Nanjing University of Information Science and Technology, Nanjing, China

2 Canadian Centre For Climate Modelling and Analysis, Environment and Climate Change Canada, University of Victoria, Victoria, British Columbia, Canada

References

[1] Lenoble J. Radiative Transfer in Scattering and Absorbing Atmospheres: Standard Computational Procedures. Hampton, VA: A. Deepak Publishing; 1985. 314 pp

[2] Toon OB, McKay CP, Ackerman TP. Rapid calculation of radiative heating rates and photodissociation rates in inhomogeneous multiple scattering atmospheres. Journal of Geophysical Research. 1989;**94**:16287-16301

[3] Fu Q, Liou KN, Cribb MC, Charlock TP, Grossman A. Multiple scattering parameterization in thermal infrared radiative transfer. Journal of the Atmospheric Sciences. 1997;**54**: 2799-2812

[4] Li J, Dobbie S, Raisanen P, Min Q. Accounting for unresolved cloud in solar radiation. Quarterly Journal of the Royal Meteorological Society. 2005;**131**:1607-1629

[5] Vane D, Tourville N, Stephens G, Kanekiewicz A. New observations of hurricanes from the cloudsat radar. In: AGU Fall Meeting Abstracts. 2006:A13A-0885

[6] Boutle IA, Abel SJ, Hill PG, Morcrette CJ. Spatial variability of liquid cloud and rain: Observations and microphysical effects. Quarterly Journal of the Royal Meteorological Society. 2014;**140**:583-594

[7] Young AH, Bates J, Curry J. Application of cloud vertical structure from cloudsat to investigate modis-derived cloud properties of cirriform, anvil and deep convective clouds. Journal of Geophysical Research. 2013;**118**:4689-4699

[8] Luo ZJ, Jeyaratnam J, Iwasaki S, et al. Convective vertical velocity and cloud internal vertical structure: An a-train perspective. Geophysical Research Letters. 2014;**41**:723-729

[9] Chen MN, Lu CS, Liu YG. Variation in entrainment rate and relationship with cloud microphysical properties on the scale of 5m. Scientific Bulletin. 2015;**60**:707-717

[10] Li J, Geldart DJW, Chylek P. Solar radiative transfer in clouds with vertical internal homogeneity. Journal of the Atmospheric Sciences. 1994;**51**:2542-2552

[11] Liou KN. An Introduction to Atmospheric Radiation. 3d ed. USA: Academic Press; 2002. 583 pp

[12] Mackel A, Mitchell DL, Bremen LV. Monte Carlo radiative transfer calculations for inhomogeneous mixed phase clouds. Physics and Chemistry of the Earth, Part B. 1998;**24**:37-241

[13] von Salzen K et al. The Canadian fourth generation atmospheric global climate model (CANAM4). Part I: Representation of physical processes. Atmosphere-Ocean. 2013;**51**: 104-125

[14] Kato T. Perturbation Theory of Linear Operator. Germany: Spring-Verlag; 1966. 18 p

[15] Meador WE, Weaver RE. Two-stream approximation to radiative transfer in planetary atmospheres: A unified description of existing methods and a new improvement. Journal of the Atmospheric Sciences. 1988;**37**:630-643

[16] Zhang F, Jia-Ren Yan J, Li K, Wu H, Iwabuchi, Yi-Ning Shi. A new radiative transfer method for solar radiation in a vertically internally inhomogeneous medium. Journal of the Atmospheric Sciences. 2018;**75**:41-55

[17] Elsasser WM. Heat Transfer by Infrared Radiation in the Atmosphere. USA: Harvard University Press; 1942. 107 p

[18] Shi Yi-Ning F, Zhang, Jia-Ren Yan H, Iwabuchi, Zhen Wang. The standard perturbation method for infrared radiative transfer in a vertically internally inhomogeneous scattering medium. Journal of Quantitative Spectroscopy and Radiative Transfer. 2018;**213**:149-158

[19] Husi L, Nakajiama TY, Matsui TN. Development of an ice crystal scattering database for the global change observation mission/second generation global imager satellite mission: Investigating the refractive index grid system and potential retrieval error. Applied Optics. 2012;**51**:6172-6178

[20] Husi L, Ishimoto H, Jerome R, et al. Investigation of ice particle habits to be used for ice cloud remote sensing for the gcom-c satellite mission. Atmospheric Chemistry and Physics. 2016;**8**:4787-4798

[21] Yang P, Liou K, Bi L, Liu C, Yi B, Baum B. On the radiative properties of ice clouds: Light scattering, remote sensing and radiation parameterization. Advances in Atmospheric Sciences. 2015;**32**

[22] Noonkester VR. Droplet spectra observed in marine stratus cloud layers. Journal of the Atmospheric Sciences. 1984;**41**:829-845

[23] Chen R, Wood R, Li Z, et al. Studying the vertical variation of cloud droplet effective radius using ship and space-borne remote sensing data. Journal of Geophysical Research. 2008;**113**:762-770

[24] Fu Q. Parameterization of radiative processes in vertical nonhomogeneous multiple scattering atmospheres [PhD thesis]. University of Utah; 1991. 259 pp

Edited by İlkay Bakırtaş

The governing equations of mathematical, chemical, biological, mechanical and economical models are often nonlinear and too complex to be solved analytically. Perturbation theory provides effective tools for obtaining approximate analytical solutions to a wide variety of such nonlinear problems, which may include differential or difference equations. In this book, we aim to present the recent developments and applications of the perturbation theory for treating problems in applied mathematics, physics and engineering. The eight chapters cover a variety of topics related to perturbation methods. The book is intended to draw attention of researchers and scientist in academia and industry.

Published in London, UK

IntechOpen

ISBN 978-1-83881-668-1