A Review on Evaluation Methods of Climate Modeling

ZHAO Zong-Ci¹,², LUO Yong¹,², HUANG Jian-Bin¹

¹Center for Earth System Science, Tsinghua University, Beijing 100084, China
²National Climate Center, China Meteorological Administration, Beijing 100081, China

Abstract

There is scientific progress in the evaluation methods of recent Earth system models (ESMs). Methods range from single variable to multi-variables, multi-processes, multi-phenomena quantitative evaluations in five layers (spheres) of the Earth system, from climatic mean assessment to climate change (such as trends, periodicity, interdecadal variability), extreme values, abnormal characters and quantitative evaluations of phenomena, from qualitative assessment to quantitative calculation of reliability and uncertainty for model simulations. Researchers started considering independence and similarity between models in multi-model use, as well as the quantitative evaluation of climate prediction and projection effect and the quantitative uncertainty contribution analysis. In this manuscript, the simulations and projections by both CMIP5 and CMIP3 that have been published after 2007 are reviewed and summarized.

Keywords: Earth system models; evaluation methods (metrics); quantitative evaluations; review

1 Introduction

In recent years, Earth system models (ESMs) and their simulations as well as projections have made progress. The coupled model intercomparison project by the WCRP has entered phase 5 (CMIP5). Two major and key international conferences on climate modeling have been held, i.e., the Open Science Conference by the WCRP in Denver (USA) in October 2011 and the Workshop on CMIP5 in Hawaii (USA) in March 2012, respectively. These conferences focused on exhibiting the further developments of the studies of the Earth system models and their simulations and projections.

Firstly, the CMIP5 collected more than 50 model experiments based on 23 model groups. These models have high resolutions with horizontally $0.5^\circ \times 4.0^\circ$ for the atmospheric component and $0.2^\circ \times 2.0^\circ$ for the ocean component. The models include more biogeochemical processes than before, design the carbon cycle and the sulfur cycle between the atmosphere and land and ocean, improve aerosol models and dynamic vegetation models, consider further dynamic and rheology in the ice models, and so on. Compared with the CMIP3, CMIP5 experiments show obvious changes, such as adding near-term climate projections and predictability, i.e., decadal predictions and projections for the future 30 years (2006–2035) based on the global coupled atmospheric and oceanic models in the Earth system models with high resolutions; the historical experiments of climate change for 1850–2005; the projections of climate change by using the new representative concentration pathways (RCPs) for the next 50–100 years or longer (up to 2300). CMIP5 required the model groups to provide simulations and projections of more variables for the annual, seasonal
and monthly time scales, also daily and hourly values than CMIP3. Therefore, it is much more useful and convenient for the users than before to investigate climate change and impacts [Taylor et al., 2012; Zhao, 2009]. Secondly, the CMIP5 prominently highlights the quantitative test method, and emphasizes the model performance metrics (quantitative measures) for the model simulations and projections of climate change. The quantitative measures refer to: 1) simulations and projections for the multi-variables at all spheres, multi-index, multi-process, multi-time scales and multi-space scales, 2) using the methods of mathematical statistics and significant tests to determine the climate model consistent at the high, medium and low levels, 3) providing evaluation evidence, which belongs to the robust, moderate or limited (ambiguity, question, very little evidence), 4) quantitative assessment of the simulated capabilities not only for climate mean state, but also for the climate change trends and periodic variations, as well as decadal and other variabilities, phenomena and extremes, 5) quantitative analysis of independence and correlation among a large number of models, 6) quantitative reliability and uncertainty for climate predictions and projections of global and various spatial scales [Taylor et al., 2012; Gleckler et al., 2008].

In the spring of 2011, the CMIP5 data center started to release the simulation results of the Earth system models. Up to the beginning of 2012, most models’ results had been released by the data center. In this review, the simulations and projections by both CMIP5 and CMIP3 (released by 2004) that have been published after 2007 are reviewed and summarized.

2 Evaluation from single to multi-variables

Since the IPCC AR4 was published in 2007, more studies have focused on the evaluation and tests of multi-variables than single variable. Among them, an interesting research result was that 24 meteorological variables simulated by 22 global climate models of CMIP3 were evaluated for the climate mean state of 1980–1999 (Fig. 1). The relative mean square errors in both space and time were calculated. The references were reanalysis data. The relative mean square error was the percentage of difference in the mean square error between each model and the typical model (i.e., medium of 22 models), respectively. For example, −0.2 for a model variable indicated a simulation better by 20% than the typical model. The number 0.2 expressed a simulation worse by 20% than the typical model. It means that the negative values indicate better simulations [Gleckler et al., 2008]. It is noticed in Figure 1 that in general, the ensemble and medium of 22 models for 24 meteorological variable simulations had better levels than each model. It is also noted that poor simulations had been found for some models and variables. It should be emphasized that the results might be quite different by using various re-analysis data as references.

The above calculation and research methods have been widely applied in the quantitative evaluation of model simulation results. For example, a recent research focused on 32 extreme variables (such as cold/warm length, cold/warm days, cold/warm nights, maximum/minimum temperature, frost/frost damage days, growth season length, daily temperature range, continuous drought/wet days, 1-day/5-day maximum precipitation, intensity of daily precipitation, precipitation in the wet period of a year, extremely/very wet days, very intensive/intensive rainfall days) simulated by 8 Earth system models of CMIP5. The evaluation methods used as the above metrics got similar conclusions, i.e., the ensemble and medium of 8 models for 32 extreme variables have better simulations than each model, as well as the results depend on various reanalysis data (Emori, personal communication). Another new study used 28 Earth system models of the CMIP5 to evaluate the simulation capability for the land surface’s monthly cold/warm extremes in the globe and 21 sub-regions. The conclusions pointed out similar results, i.e., the ensemble and medium of 28 models of CMIP5 have better simulations than each model. It is also found that the evaluations for the ensemble and medium of 28 models of CMIP5 and 24 models of CMIP3 had similar simulations with the high levels, respectively. It was also found that CMIP5 did not have better simulations than CMIP3.
Figure 1 Portraits diagram of relative error metrics for CMIP3 global annual cycle climatology (1980–1999). A diagonal splits each grid square in order to show the relative error with respect to both ERA40 (upper left triangle) and NCEP/NCAR (lower right triangle) reanalysis data. The horizontal ordinate are 22 CMIP3 models (from left to right: model ensemble mean, model ensemble median, Norway model, two Canada models, France model, Australia model, five US models, China model, Russia model, France model, two Japan models, Korea/Germany model, Germany model, Japan model, two US models, two UK models, respectively), the vertical ordinate are 24 meteorological variables (from top to bottom: surface latent flux, surface sensible flux, surface temperature, reflect solar radiation in clear sky, reflect solar radiation, outgoing long wave radiation in clear sky, outgoing long wave radiation, total cloud cover, precipitation, water vapor in a column, sea level pressure, meridional wind stress, zonal wind stress, surface meridional wind, surface zonal wind, absolute humidity at 400 and 850 hPa, meridional wind, zonal wind and temperature at 850 hPa, respectively) \cite{Gleckler et al. 2008} (Yao, personal communication).

3 Evaluation from climate mean state to climate change and attributions

In earlier assessments, the climate model evaluation focused on simulation capability of climate mean state compared with the observation or reanalysis data. In the past 10 years, most evaluations concentrated on climatic changes, such as climate trends, periodicity or decadal variabilities, as well as attributions, besides climate mean state. In recent years, climate trends have been investigated by using the CMIP5 models. Table 1 and Figure 2 present the evolutions of the annual global mean surface air temperature simulated by the Canadian Earth system model (CanESMs) of CMIP5 with the full forcing (including both natural and anthropogenic forcings), only natural forcing, only anthropogenic forcing, and only aerosol forcing for 1851–2010. Five runs and their ensemble have been calculated. In the three selected stages, the simulated climate trends for the full forcing were very close to the observed trends (Table 1) and the evolutions of the annual mean global average temperature
Figure 2. Time series of global mean surface air temperature anomalies in observations and simulations of CanESM2 of CMIP5. Black lines show observed global mean annual mean temperatures from HadCRUT3, and the thin colored lines show the global mean temperature from five-member ensembles of CanESM2 forced with (a) anthropogenic and natural forcings (ALL), (b) natural forcing only (NAT), (c) greenhouse gases only (GHGs), and (d) aerosols only (AER). All anomalies are calculated relative to the period 1851–1900, and ensemble means are shown by the thick colored lines [Gillett et al., 2012].

were similar to the observation (Fig. 2). A study on some other Earth system models of CMIP5 also confirmed these conclusions [Gillett et al., 2012].

A recent study included 42 simulation experiments by using 6 Earth system models of CMIP5 and compared these with two reanalysis data sets. It detected warming trends of 0.32°C per decade by the models slightly lower than the observation (0.49°C per decade). It is found that the obvious warming of 0.39, 0.38 and 0.49°C per decade in the area between 20°N and the North Pole occurred in spring, summer and fall seasons of the Northern Hemisphere, respectively, but slight warming of 0.07°C per decade in winter only. It means that seasonal warming was asymmetric over the extratropical areas of the Northern Hemisphere. In fact, in recent years, the global warming amplitude was decreasing mainly due to significantly reduced warming in boreal winter. Winter warming reduced or even became cold mainly located in the northern part of North America and Europe, as well as Siberia. The CMIP5 models were able to simulate the warming trends in spring, summer and fall seasons north of 20°N in the Northern Hemisphere, but they had difficulties to simulate the slight warming or cooling in wintertime in the same regions.

Table 1. Attributable temperature changes trends during three periods for greenhouse gases (GHGs), other anthropogenic forcing (OTH), natural forcing (NAT) and observations (OBS) (based on Gillett et al. [2012]; unit: °C)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>OBS</td>
<td>0.60</td>
<td>0.50</td>
<td>0.75</td>
</tr>
<tr>
<td>GHGs</td>
<td>0.75</td>
<td>0.70</td>
<td>0.80</td>
</tr>
<tr>
<td>OTH</td>
<td>-0.15</td>
<td>-0.10</td>
<td>-0.03</td>
</tr>
<tr>
<td>NAT</td>
<td>-0.05</td>
<td>-0.15</td>
<td>0.03</td>
</tr>
<tr>
<td>GHGs+OTH+NAT</td>
<td>0.55</td>
<td>0.45</td>
<td>0.80</td>
</tr>
</tbody>
</table>
4 Analysis of uncertainty contributions from qualitative assessments to quantitative evaluation of projections

Climate models are often selected as instruments to project climate change for the next 50–100 years under the increasing of anthropogenic GHG and aerosol emissions. However, the climate projections by models have significant uncertainties, especially on regional scale. Therefore, the contributions of projection uncertainties should be calculated quantitatively. In recent years, a study which calculated uncertainty contributions of changes in annual mean global average surface air temperatures in the 21st century by 15 climate models of CMIP3 [Pirani et al., 2009] has attracted attention. It shows that considering the uncertainty contributions of climate models, emission scenarios and internal variability (unforced variability), in the early 21st century, climate models had the largest contribution to the uncertainty, and later, the largest uncertainty contribution was from the emission scenarios. For instance, in the 2010s, the uncertainty contributions of the projected global temperature variation from climate models accounted for about 65%, from the emission scenarios about 2%, and from internal variability about 33%, respectively. In the 2090s, the main uncertainty contribution was from the emission scenarios (81%), but from climate models only 17% and from internal variability about 2%, respectively. There were a few differences in the uncertainty contributions between the projected global and regional temperatures. Taking China as an example, in the 2010s, the uncertainty contributions from climate models were about 48%, from internal variability about 42%, from emission scenarios about 10%, respectively. In the 2090s, they were about 70% from emission scenarios, 25% from climate models and 5% from internal variability, respectively. Different regions show slightly different results, but they are generally consistent [Pirani et al., 2009; Zhao et al., 2009]. The conclusions are that the uncertainty from climate models is important, especially for the early decades of the 21st century. However, the anthropogenic emission scenarios become the main uncertainty contributor for the late decades of the 21st century.

A new study on the uncertainty contributions projected the changes of sea surface temperature (SST) and tropical storm frequency in the North Atlantic Ocean by using 17 Earth system models of CMIP5 with RCP2.6, RCP4.5 and RCP8.5, respectively. The Earth system models contributed to the main uncertainty for the projected SST in the early decades of the 21st century, but the emission scenarios (RCPs) are the significant contributor in the late decades of the 21st century. It should be emphasized that the uncertainty of the Earth system models always played a leading role for the projected tropical storm frequency, and the contribution from internal variability was also obvious, especially in the early decades of the 21st century. The uncertainty of emission scenarios (RCPs) had a small contribution to tropical storm frequency (Table 2) [Villarini and Vecchi, 2012].

5 Evaluation from climate model agreement to model independence

In the reliability study, the evidence (robust, medium and limited) and agreement (high, medium and low) of the confidence level is one standard of credible degrees for quantitatively assessing the results. Among them, the agreement numbers of the model results are often used to estimate the statistical confidence levels of the simulations and projections. The more consistent the models’ results, the greater
Table 2 Fractional contribution (in %) to uncertainties of SST and tropical storm frequency over the Northern Atlantic Ocean as projected by 17 CMIP5 Earth system models (calculated on Villarini and Vecchi [2012])

<table>
<thead>
<tr>
<th>Projection</th>
<th>CMIP5 Scenario Internal variability</th>
<th>CMIP5 Scenario Internal variability</th>
</tr>
</thead>
<tbody>
<tr>
<td>SST</td>
<td>85 5 15 20 78 2</td>
<td>2080 80 15 20 78 0 2</td>
</tr>
<tr>
<td>Tropical storm</td>
<td>50 1 49 75 0 25</td>
<td></td>
</tr>
</tbody>
</table>

the reliability. In another way, generally speaking, the multi-model ensembles adopt the ensemble mathematical means. It means that each model has the equal weight, i.e., one model, one vote. However, many models have same or similar parameterization schemes or belong to the same class. It causes relevance and dependence between the climate models. Therefore, some recent studies concentrate on the independence and relevance of the models.

A recent study attempted to provide the model family tree (classification and ascription) for surface air temperature and precipitation based on the simulations of climate mean states, seasonal cycles, interannual variability and space correlations by the control runs of 24 climate models of CMIP3. The reanalysis data (ERA, NCEP and MERRA for temperature, and GPCP and CMAP for precipitation) were also added to the model family tree (Fig. 4). It is found in Figure 4 that some models among the 24 models belonged to the same or similar classes. The author pointed out that the relevance and dependence of the climate models might provide false evidence when one estimated the model agreement numbers and calculated the model ensembles. Therefore, model independences must be considered [Masson and Knutti, 2011]. Another study selected different classification standard, i.e., the biases from 35 climate variables between observations and simulations by CMIP3 models. Similar results have been noticed [Pennell and Reichler, 2011]. Summarizing the above mentioned, as considering the agreement numbers of simulations and projections by climate models, one should be careful to think about model relevance and similarity, which might cause the distortion of reliability and confidence levels for simulations and projections by climate models.

![Figure 4](link-to-figure)

Figure 4 Hierarchical clustering of 24 CMIP3 models for surface air temperature (a) and precipitation (b) in the model control state. Models from the same institution or sharing versions of the same atmospheric models are shown in the same color. Observations are also marked in black [Masson and Knutti, 2011]
6 Evaluation from single model for a variable to multi-models for processes, phenomena and extremes

For recent years, more and more studies evaluated the reliability of climate processes, phenomena and extremes simulated and projected by climate models. A new study evaluated the global monsoon, such as the global monsoon areas (annual precipitation intensity over 2 mm d\(^{-1}\), and summer precipitation over 55% to the annual total), the global monsoon total precipitation and the global monsoon intensity (the global monsoon precipitation per unit area) projected by 24 models of CMIP3 with SRES A1B scenario and 11 Earth system models of CMIP5 with RCP4.5 for 2075–2099 relative to 1979–2003, respectively. Both the CMIP3 and CMIP5 project the increasing global monsoon areas, total precipitation and intensity under global warming (Table 3). On the other hand, five global atmospheric models with high resolutions (2 German MPI models with a horizontal resolution of 1.125\(^\circ\); 1 German MPI model with a horizontal resolution of 40 km; 1 Japanese MRI model with a horizontal resolution of 20 km, 1 USA GFDL model with a horizontal resolution of 50 km) driven by the projected SST, projected to enhance three global monsoon indices mentioned above. That is, five global atmospheric models with high resolutions, 24 CMIP3 models and 11 CMIP5 models provided the coincident projections on the future strengthening global monsoon [Hsu et al., 2012].

7 From simple predictions and projections to quantitative evaluation of their effects

As known, projections of future climate change by using global climate models under various emission scenarios have been conducted for about 20 years. The IPCC AR4 has evaluated quantitatively the global temperature changes projected by the climate models with various emission scenarios. As an example in IPCC AR4, Table 4 gives the global mean annual average temperature anomalies for 2001–2011 relative to 1961–1990 projected by 23 CMIP3 models with four scenarios (SRES A2, A1B and B1, as well as commitment at year 2000 level) [IPCC, 2007]. It can be seen from Table 4 that the projected annual mean temperature anomalies under four scenarios were positive in each year of 2001–2011 relative to 1961–1990. Compared with the observations, the models with the four emission scenarios succeeded to project the warming characteristics for the first 11 years of the 21st century. The problem remains that the models projected obvious warming trends that were not shown by the observations, especially for the last six years in which the projected warming was much larger than observations. The reason should be investigated in future.

8 Conclusions and discussion

This paper reviewed the evaluation methods of the Earth system models and obtained the following
conclusions:

(1) The integrated evaluations for the Earth system models should concentrate on the multi-variables in five spheres; it is not enough to evaluate only a single variable.

(2) Climate change (such as trends, periodicity and decadal variability) simulated and projected by the models must be evaluated quantitatively, besides the climate mean states.

(3) The uncertainty contributions of many factors on climate changes projected by the models should be evaluated quantitatively, not only qualitatively.

(4) In considering the agreements of simulations of the models, we should not focus on the model results merely, but also notice the similarity and dependence among the climate models in order to achieve the independence of the evidence.

(5) There is need not only to quantitatively evaluate the simulations of single variables, but also to evaluate the simulations of the processes, phenomena and extremes quantitatively in the five spheres.

(6) The evaluations for predictions and projections of future climate change must be quantitative. The best value with a range at certain confidence level should be provided.

The advances of evaluation methods for the Earth system models depend on a variety of factors, such as the quality of observed data, the knowledge of the physical, chemical and biological processes in the five spheres of the Earth system, the development of mathematical methods of evaluation and their application, as well as the computer technique. Currently, the IPCC AR5 report is being produced. Many new studies and evaluation methods are expected, such as those on quantitatively testing the simulation capability of climate change in the five spheres by the Earth system models and regional climate models, quantitative attribution analysis, decadal and near-term predictions as well as long-term projections, and uncertainty contribution analysis.

Acknowledgements

This paper was supported by the Ministry of Science and Technology 973 Project (No. 2010CB950501-03), and the National Natural Science Foundation (No. 41175066).

References


