https://doi.org/10.5194/gmd-2024-10
Preprint. Discussion started: 29 January 2024

Geoscientific Model Development Discussions

A Deep Learning-Based Consistency Test Approach for Earth System Models on Heterogeneous Many-Core Systems

Yangyang Yu[1], Shaoqing Zhang[*1,2,3], Haohuan Fu[*4,5], Dexun Chen[*5], Yang Gao[3,6], Xiaopei Lin[1,2,3], Zhao Liu[5,7], Xiaojing Lv[5,8]

[1] Key Laboratory of Physical Oceanography, Ministry of Education, Ocean University of China, Qingdao, 266100, China
[2] Frontiers Science Center for Deep Ocean Multispheres and Earth System (FDOMES), Academy of the Future Ocean, College of Oceanic and Atmospheric Sciences, Ocean University of China, Qingdao, 266100, China
[3] Laoshan Laboratory, Qingdao, 266100, China
[4] Ministry of Education Key Lab. for Earth System Modeling, and Department of Earth System Science, Tsinghua University, Beijing, 100084, China
[5] National Supercomputing Center in Wuxi, Wuxi, 214072, China
[6] Key Laboratory of Marine Environmental Science and Ecology, Ministry of Education, Frontiers Science Center for Deep Ocean Multispheres and Earth System (FDOMES), Ocean University of China, Qingdao, 266100, China
[7] Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, China
[8] China Ship Scientific Research Center, Wuxi, 214082, China

Correspondence to: Shaoqing Zhang (szhang@ouc.edu.cn), Haohuan Fu (haohuan@tsinghua.edu.cn), Dexun Chen (adch@263.net)

Abstract. Physical and heat limits of the semiconductor technology require the adaptation of heterogeneous architectures in supercomputers to maintain a continuous increase of computing performance. The coexistence of general-purpose cores and accelerator cores, which usually employ different hardware architectures, can lead to bit-level differences, especially when we try to maximize the performance on both kinds of cores. Such differences further lead to unavoidable computational perturbations through temporal integration, which can blend with software or human errors. Software correctness verification in the form of quality assurance is a critically important step in the development and optimization of Earth system models (ESMs) on heterogeneous many-core systems with mixed perturbations of software changes and hardware updates. We have developed a deep learning-based consistency test approach for Earth System Models referred to as ESM-DCT. The ESM-DCT is based on the unsupervised bidirectional gate recurrent unit-autoencoder (BGRU-AE) model, which can still detect the existence of software or human errors when taking hardware-related perturbations into account. We use the Community Earth System Model (CESM) on the new Sunway system as an example of large-scale ESMs to evaluate the ESM-DCT. The results show that facing with the mixed perturbations caused by hardware designs and software changes in heterogeneous computing, the ESM-DCT can detect software or human errors when determining whether or not the model simulation is consistent with the original results in homogeneous computing. Our ESM-DCT tool provides an efficient and objective approach for verifying the reliability of the development and optimization of scientific computing models on the heterogeneous many-core systems.

https://doi.org/10.5194/gmd-2024-10
Preprint. Discussion started: 29 January 2024

1 Introduction

The improvement of resolution and complexity of the numerical models requires the increase in computing power. As the Moore's Law slows down (Kish, 2002), the increase in the number of processors, the design of inexact hardware (Düben et al., 2014), and the greener architecture of heterogeneous many-core (Fu et al., 2016) are used to continue increasing the computing performance. For greener and faster heterogeneous many-core architectures, the major computing power is provided by many-core accelerators such as NVIDIA graphics processing units (GPUs) (Vazhkudai et al., 2018) as well as Intel Xeon Phi MICs (Liao et al., 2014) and many-core processors Sunway computing processing elements (CPEs) (Fu et al., 2016).

Earth system models (ESMs) are based on mathematical equations and established by dynamical, physical, chemical, and biological processes through numerical methods consisting of millions of lines of legacy codes (Flato, 2011), such as the Community Earth System Model (CESM; Hurrell et al., 2013). Science advancement and societal needs require ESMs with higher and higher resolution to resolve more details of interacting atmosphere, ocean, sea-ice, and land surface components, which demands tremendous computing power. Therefore, ESMs are the important application scenarios for the heterogeneous many-core high-performance computing (HPC) systems (Gu et al., 2022; Zhang et al., 2023).

However, there are the differences of hardware design between the general-purpose cores and accelerator cores in the heterogeneous many-core architectures. Compared with homogeneous computing using the general-purpose cores only, heterogeneous computing can cause nonidentical floating-point outputs whenever a accelerator processor or accelerator is involved (Yu et al., 2022). The computational perturbations caused by hardware designs can blend with software or human errors, which can affect the accuracy of the model verification (Zhang et al., 2020).

Model verification during optimizing and developing ESMs is critical to establishing and maintaining the credibility of the ESMs (Carson II, 2002), which focuses on determining whether or not the implementation of a model is correct and matches the intended description and assumptions for the model. Evaluating the scientific consistency is a commonly used method for model verification in the form of quality assurance. For example, for detecting the influences of hardware environment changes, data from a model simulation of several hundred years (typically 400) on the new machine is analyzed and compared to data from the same simulation on a trusted machine by climate scientists (Baker et al., 2015). Then, the CESM ensemble-based consistency test (CESM-ECT) is used to compare the new simulations against the control ensemble from the trusted machine (Baker et al., 2015; Milroy et al., 2016; Baker et al., 2016; Milroy et al., 2018). However, all the methods mentioned above focus on homogeneous multi-core HPC systems. Facing the scenario of mixed perturbations composed of the inevitable computational perturbations caused by hardware designs (Yu et al., 2022) and software/human errors, there is an urgent demand to evaluate the scientific consistency for the ESMs on the heterogeneous HPC systems, which can accept

https://doi.org/10.5194/gmd-2024-10
Preprint. Discussion started: 29 January 2024


Geoscientific Model Development Discussions

the influence of the computational perturbations for further detection of software or human errors generated in optimizing and developing the ESMs.

With the improvement of resolution and complexity of ESMs, there exists an urgent demand on evaluating consistency rapidly. Milroy et al. have reduced the length of the ensemble to save the potential further cost while improving the CESM-ECT. However, the ultra-fast tests in the CESM-ECT are applied for evaluating the scientific consistency on the Community Atmosphere Model (CAM; Milroy et al., 2018). The CESM-ECT for ocean component still requires simulations of one year or more. There is a lack of a method to analyze short-time simulation results of multi-components, which can achieve an overall consistency evaluation of ESM rapidly. Besides, the principal component analysis (PCA) in the CESM-ECT is just applied to exploring linear patterns contained in the confusing datasets (Liu et al., 2009). Facing with the non-linear relationship generated by the combination of multi-component data, there is an urgent need of data analysis methods for non-linear transformation, such as deep learning models.

The goal of this article is to develop a deep learning-based consistency test approach for the ESMs, referred to as ESM-DCT, on the heterogeneous many-core HPC system. The ESM-DCT tool is based on the unsupervised bidirectional gate recurrent unit-autoencoder (BGRU-AE; Zhao et al., 2017) model, which can accept the unavoidable computational perturbations caused by hardware designs. The ESM-DCT is applied to evaluate whether or not a new CESM configuration in the scenario of mixed perturbations composed of the inevitable computational perturbations and software or human errors in the heterogeneous computing is consistent with the original “trusted” configuration in the homogeneous computing.

The rest of the paper is organized as follows. Section 2 shows additional background information. Section 3 details the deep learning-based consistency test approach. Section 4 shows the results of experiments with the consistency test approach. Finally, the summary and discussion are given in Section 5.

2 Background

The new Sunway system is the new generation of Chinese home-grown supercomputer that inherits and develops the architecture of Sunway TaihuLight (Fu et al., 2016). The new Sunway system is built using an upgraded heterogeneous many-core processor, SW26010P, which is similar to SW26010 in terms of architecture but with more computing cores and higher overall HPC capability. Each SW26010P processor can be divided into six identical core groups (CGs), which are connected through the network on the chip, as shown in **Figure 1**. Each CG includes one management processing element (MPE) and one computing processing element (CPE) cluster with 8x8 CPEs (Gu et al., 2022). Each CPE has its own instruction cache and data storage that can be configured as a fully user-controlled local data memory (LDM) or can be configured as a partly automatically hardware-managed local data cache (LD cache).

https://doi.org/10.5194/gmd-2024-10
Preprint. Discussion started: 29 January 2024

There are the hardware differences between the general-purpose cores and accelerator cores in Sunway systems. This architectural distinction is designed to optimize aggregated computing power while minimizing micro-architectural complexities. MPEs and CPEs serve distinct functions, promoting a hybrid computational approach that utilizes separate instruction sets (Fu et al., 2016; Fu et al., 2017a; Fu et al., 2017b), thus leading to hardware-generated differences in corner cases related to denormalized numbers. Therefore, there can exist unavoidable computational perturbations caused by the hardware design (Yu et al., 2022), which should be accepted for further detection of software or human errors generated in optimizing and developing the ESMs. The key challenge is designing a tool to evaluate the scientific consistency, which can remove the influences of heterogeneous perturbations.

Coarse-grained testing is a common practice for software verification in the form of quality assurance (Clune and Rood, 2011). Coarse-grained testing does not offer information as to the source of the inconsistency but rather as to whether or not the inconsistency may exist (Baker et al., 2015). This coarse-grained testing typically takes the form of analyses of simulation results (Easterbrook and Johns, 2009) and issues an overall pass or fail result, such as CESM-ECT (Baker et al., 2015). However, the principal component analysis (PCA) in the CESM-ECT is just applied to exploring linear patterns contained in the confusing datasets (Liu et al., 2009). Facing with the non-linear relationship generated by the combination of multi-component data in the coupled numerical model, there can use deep learning models to handle data by embedding multiple non-linear activation functions. In the last several years, the unsupervised deep learning model has been a fresh and widely used data mining method and explored for coarse-grained anomaly detection. Such deep learning method can efficiently and objectively identify whether or not the data are different from the original status (Fotiadou et al., 2021; Maya et al., 2019). Hence our focus in this work is on an unsupervised deep learning-based tool for coarse-grained testing to detect the consistency in the features of simulation results on the trusted and heterogeneous HPC systems.

In this study, we show the performance of deep learning models in mining non-linear features from coupled models. We design a experiment which uses the PCA model of the CESM-ECT, referred to as ECT, and a experiment which uses the BGRU-AE deep learning approach to establish a consistency test approach, referred to as DCT, to the evaluate the consistency of the 5-variable conceptual coupled model (5VCCM; Zhang, 2011) simulations. The 5VCCM is a conceptual atmosphere-ocean coupled model that essentially couples the 3-variable Lorenz63 model (Lorenz 1963) to a slab-ocean variable and a simple pycnocline predictive model (Gnanadesikan 1999). The governing equations are given by Eq.(1):

$$\begin{aligned}
\dot{\chi}_1 &= -\sigma\chi_1 + \sigma\chi_2 \\
\dot{\chi}_2 &= -\chi_1\chi_3 + (1 + C_1\omega)k\chi_1 - \chi_2 \\
\dot{\chi}_3 &= \chi_1\chi_2 - b\chi_2 \\
O_m\dot{\omega} &= C_2\chi_2 + C_3\eta + C_4\omega\eta - O_d\omega + S_m + S_s\cos(2\pi t / S_{pd}) \\
\Gamma\dot{\eta} &= C_5\omega + C_6\omega\eta - O_d\eta
\end{aligned}, \qquad (1)$$

https://doi.org/10.5194/gmd-2024-10
Preprint. Discussion started: 29 January 2024


Geoscientific Model Development Discussions

where χ_1, χ_2, and χ_3 denote the atmospheric model variables; ω and η represent the upper and deep ocean states, respectively. Within the atmospheric component, the standard values of σ, k and b are set as 9.95, 28 and 8/3, respectively, to maintain the chaotic atmospheric nature. In the upper ocean equation, the parameters O_m and O_d represent the heat capacity and damping coefficients of the ocean, which are set as 1 and 10, respectively. The term $S_m + S_s\cos(2\pi t / S_{pd})$ denotes the external forcing, where the parameters S_m and S_s represent the annual mean and seasonal cycles of external forcing and are set as 10 and 1, respectively. The parameter S_{pd}, representing the model's seasonal cycle, is set as 10 to ensure that the external forcing period is comparable to the time scale of the upper ocean. In the equation for the deep ocean component, the ratio of Γ to O_d defines its time scale, which is longer than that of the upper ocean component (the ocean slab's time scale is 1/10 that in the deep ocean; therefore, Γ is defined as 100). The coupling coefficients C_1 and C_2 included in the χ_2 and ω equation simply denote the coupling interaction between the oceanic slab and atmospheric component and are defined as 0.1 and 1, respectively, with C_1 representing the upper oceanic forcing compared to the atmospheric component and C_2 representing the opposite relationship. Similarly, the parameters C_3, C_4, C_5 and C_6 are used to implement the interference between the deep and slab oceans and are respectively set as 0.01, 0.01, 1 and 0.001, with C_3 denoting deep oceanic forcing onto the slab ocean and C_5 denoting the opposite, and C_4 and C_6 representing nonlinear interaction. All variables are non-dimensional (Zhao et al., 2019). A fourth-order Runge-Kutta time difference scheme and a leap-frog timedifference scheme with a Robert-Asselin time filter (Asselin 1972), are used to resolve this 5VCCM, with the time step equaling 0.01 TU (time unit) (1 TU=100 time steps).

We generate a 151-member ensemble of 1000 time steps in the 5VCCM in the homogeneous computing using only the MPE as the training datasets. Then, we use the method of the ECT and DCT to evaluate the 5VCCM simulations with the different compilers and heterogeneous computing, as shown in **Table 1**. The model parameters are obtained from transfer learning of CESM-ECT and ESM-DCT detailed in section 4. Note that by default, ECT evaluates 3 simulations for each test scenario and issues an overall failure (meaning the results are statistically distinguishable) if more than two of the PC scores are problematic in at least two of the test runs. The DCT evaluates 40 for each test scenario and issues an overall failure if the passing rate is less than 90%. **Table 1** shows that the 5VCCM with GNU compiler passes the ECT, but the 5VCCM with Intel compiler and heterogeneous computing fails. The 5VCCM with GNU/Intel compilers and heterogeneous computing pass the DCT. However, the modifications with GNU/Intel compilers and heterogeneous computing can lead to nonidentical floating-point results but are not expected to be scientifically-changing. The DCT shows the better performance in mining non-linear features from coupled models than the ECT, which provides the confidence in the development of a deep learning-based consistency test approach for multi-component data of ESMs on heterogeneous many-core systems.

https://doi.org/10.5194/gmd-2024-10
Preprint. Discussion started: 29 January 2024

Geoscientific Model Development Discussions

3 A Deep learning-based tool to evaluate consistency

3.1 The general idea

As noted, the unavoidable perturbations caused by the heterogeneous many-core architecture can blend with software or human errors, which can affect the accuracy of the model verification in form of evaluating the scientific consistency. Therefore, our tool must accept the unavoidable perturbations caused by hardware designs on heterogeneous HPC systems to evaluate the scientific consistency in the scenario of mixed perturbations of software changes and hardware updates. Then, we accept the new ESM configurations if the non-linear features learned from the output data can match that of the original data in the homogeneous computing. We develop a deep learning-based consistency test approach for the ESMs, referred to as ESM-DCT. Our tool includes: 1) designing an unsupervised deep learning model based on the BGRU-AE model, to study the non-linear features from the multi-component simulation results. 2) developing an ensemble approach to take the statistical distributions from short-term simulations on the multi-component as the datasets. 3) implementing the software tool to accept the unavoidable perturbations and evaluate the consistency on the heterogeneous HPC systems.

3.2 A BGRU-AE deep learning model to study the features

It is necessary to select the appropriate deep learning model based on the characteristics of the studied input data. Evaluating the consistency of simulation results should utilize as many outputs of variables as possible. Multiple variable results will be combined into sequence data, which is suitable for analysis in Recurrent Neural Network (RNN; Lukosevicius and Jaeger, 2009). Based on the RNN, Long Short-Term Memory (LSTM; Hochreiter and Schmidhuber, 1997; Song et al., 2021) introduces the "gate" in neurons to solve the problem of gradient vanishing and gradient explosion during long-term dependency processes. Then, Gate Recurrent Unit (GRU; Cho et al., 2014; Zhao et al., 2017), which is an improved LSTM, is used for fast convergence and computational cost saving. Each neuron of GRU has a reset gate r_t, which adjusts the incorporation of new input with the previous memory, and an update gate z_t, which controls the preservation of the previous memory. The reset gate r_t and the update gate z_t are defined by Eq.(2) and Eq.(3):

$$z_t = \sigma(W^z x_t + V^z h_{t-1} + b_z) \quad \text{and} \tag{2}$$

$$r_t = \sigma(W^r x_t + V^r h_{t-1} + b_r), \tag{3}$$

where W, V, and b are shared by all time steps and learned during model training, σ is the activation function. After obtaining the signal of the reset gate r_t, the GRU computes the reset result $\tilde{h}_t$, which is similar to the memory process of LSTM. The reset result $\tilde{h}_t$ is defined by Eq.(4):

$$\tilde{h}_t = \tanh(W^c x_t + V^c (r_t \otimes h_{t-1})), \tag{4}$$

https://doi.org/10.5194/gmd-2024-10
Preprint. Discussion started: 29 January 2024

Geoscientific Model Development Discussions

where W and V are shared by all time steps and learned during model training, $\otimes$ denotes the element-wise product, and tanh is the activation function. Then, the GRU updates the current hidden output h_t. The hidden output h_t is defined by Eq.(5):

$$h_t = (1 - z_t) \otimes h_{t-1} \oplus z_t \otimes \tilde{h}_t , \tag{5}$$

where $\otimes$ denotes the element-wise product. The illustration of GRU is shown in **Fig. 2**.

Traditional neural network models can only achieve forward propagation of information. However, the bidirectional models can capture the time dependencies within a sequence in a forward and backward manner (Yu et al., 2019; Su and Kuo, 2019). The basic idea of a bidirectional neural network is to input the same input sequence into the neural network that propagates forward and backward respectively. These two neural networks are connected to the same output layer, which can obtain context information. The hidden output H is defined by Eq.(6):

$$H = H_t^f \oplus H_1^b , \tag{6}$$

where H_t^f and H_1^b are the final hidden states resulting from the forward and backward processes respectively.

In deep learning, the Autoencoder (AE) is a type of neural network which can learn sequence embedding efficiently in an unsupervised manner (Hinton et al., 2006; Vincent et al., 2008). The AE consists of an encoder and a decoder. The encoder learns to compress the input data into a short code C, whereas the decoder learns to decompress the code into a set of output data O which is used to compute the loss function with input data. In this study, we propose the BGRU-AE deep learning model for consistency evaluation on the heterogeneous HPC systems. The encoder and decoder adopt the BGRU model to analyze the features. The outputs of the decoder are input to Fully Connected (FC) layers for dimensional alignment (Sun et al., 2013). The equation of FC is as follows:

$$O = f(W_{fc} * H_d + b_{fc}) , \tag{7}$$

where O is the final result of the BGRU-AE deep model, H_d is the output of the decoder, W_{fc} and b_{fc} are learned during model training. The illustration of the BGRU-AE model is shown in **Fig. 3**.

3.3 An ensemble approach to get the datasets

The development of a tool for the BGRU-AE model necessitates the consistency evaluation in the features of simulation results on the datasets. Characterizing the natural variability is difficult with a single run of the original simulation. The statistical distribution from the ensembles extended by the sampling of the original data can represent possible system states. Ensembles created by small perturbations to the initial conditions are commonly used in climate modeling to reduce the influence of the initial condition uncertainty (Sansom et al., 2013) and diagnose the influence of computing environment changes (Düben et al. 2017; Prims et al. 2019; Rosinski and Williamson, 1997; Arteaga et al., 2014). In this study, we

https://doi.org/10.5194/gmd-2024-10
Preprint. Discussion started: 29 January 2024

generate the ensemble for the datasets of consistency evaluation by running simulations that differ only in a random perturbation of the initial atmospheric temperature.

Based on the ultra-fast tests, we can analyze the short-term simulation results on the ocean and atmosphere component, to achieve the consistency evaluation for the multivariate data of CESM. We examine 97 variables from the atmosphere component results, as redundant variables and those with no variance are excluded. We examine the temperature, salinity, zonal velocities, and meridional velocities concerning the model grid from the ocean component results. In order to quickly spread the perturbations, we modify the coupling frequency of the ocean component to 8 times a day. Then, we use the simulation results at the 24 time step as the input data of the BGRU-AE model with -O2 compiling optimization options, where the ocean component is coupled 4 times.

In this study, the ensemble simulation data as the training, validation, and testing datasets are generated as needed, as listed in **Table 2**. The number of training sets, validation sets, and testing sets is 120, 40, 40. The descriptions of the datasets are as follows:

Firstly, the training datasets and validation datasets are the ensembles with the $O(10^{-14})$ perturbations of initial atmospheric temperature in the homogeneous computing using the MPE only to train the parameters of the BGRU-AE model. The BGRU-AE model is unsupervised, which needs no manual labeled data and focuses on whether or not non-linear features learned from the testing datasets can match that of the training datasets in the homogeneous computing.

Secondly, we focus on the modifications that lead to non-bit-for-bit results but are not expected to be climate-changing. The modifications include the heterogeneous programming and -O0/-O1 compiling optimization options. These testing datasets are with the $O(10^{-14})$ perturbations of initial atmospheric temperature. We expect that the modifications will be consistent with our initial ensemble distribution, which can accept the hardware-related perturbations and low-level optimizations in the heterogeneous computing.

Thirdly, our tool must successfully detect modifications to the simulation results that are known to be climate-changing. The modifications include unacceptable initial perturbations and unacceptable CESM model parameter adjustments listed by climate scientists (Baker, et al., 2015). We add $O(10^{-7})$, $O(10^{-6})$, and $O(10^{-5})$ perturbations of initial atmospheric temperature. The probability density function (PDF) of the CESM at the 24 time step is shown in **Fig. 4**. Therefore, the testing datasets tagged as unacceptable initial perturbations are the ensembles with the $O(10^{-6})$ perturbations of initial atmospheric temperature, whose PDF is clearly inconsistent with that of the $O(10^{-14})$ initial perturbations. The testing datasets with unacceptable CESM model parameter adjustments are with the $O(10^{-14})$ perturbations of initial atmospheric temperature.

Finally, we show the results for modifications with unknown outcomes in the heterogeneous computing. The modifications include -O3 compiling optimization option, mixed precision programming, and the CESM model parameter adjustments with unknown effects. These testing datasets are with the $O(10^{-14})$ perturbations of initial atmospheric temperature.

https://doi.org/10.5194/gmd-2024-10
Preprint. Discussion started: 29 January 2024


Geoscientific Model Development Discussions

3.4 A software tool to accept the unavoidable perturbations and evaluate the consistency

We further discuss the software ESM-DCT tool for the consistency evaluation of the multi-component data in the CESM on the heterogeneous many-core HPC systems. Firstly, the ensemble simulation data as the training, validation, and testing datasets are generated and handled as needed, as noted in section 3.3. All of the datasets are the simulation results of CESM at the 24 time step where the ocean component is coupled 4 times. We examine 97 variables from the atmosphere component results and 4 variables from the ocean component results. The global area-weighted mean is calculated for each variable, which is converted to a 101 dimensional vector. Then, the vector is standardized by min-max normalization because the variables have vastly different units and magnitudes.

Secondly, the BGRU-AE model should be trained using the training datasets and validation datasets. At the each time step of training the BGRU-AE model, we input the validation datasets into the BGRU-AE model and calculate the reconstruction errors using the Mean Squared Error (MSE) function. We adjust the BGRU-AE model parameters until the minimum reconstruction errors of the validation datasets and save the final BGRU-AE model.

Thirdly, the threshold of reconstruction errors of the training datasets is calculated as an indicator to issue an overall pass or fail result. We re-input the training datasets into the saved BGRU-AE deep learning model and regain the maximum value of the reconstruction errors. It is generally assumed that the reconstruction errors can be lower for the normal input since they are close to the training datasets, while the reconstruction errors can become higher for the abnormal input (Gong et al., 2019). Therefore, if the reconstruction error of one member of the testing datasets can be higher for those of maximum value of the training datasets, then the software tools for the consistency evaluation return the member as “failure”. If the passing rate is less than 90%, the tool issues an overall “failure”, which yields the accuracy of 97.9 % in the simulations with the climate-changing and non climate-changing modifications.

Finally, the software tool is applied to evaluate the consistency for the CESM on the new Sunway system. We input the testing datasets into the saved BGRU-AE model and calculate the passing rate using the reconstruction errors. We focus on the influence of the non-climate changing, climate changing, and unknown outcomes modifications on the consistency of simulation results in the heterogeneous computing. The tool can detect the existence of software or human errors when taking hardware-related perturbations into account. **Figure 5** illustrates the workflow for the software tool.

4 Experimental studies

4.1 Training the BGRU-AE model

The datasets are the simulation results of the 1.3 release series of CESM using a present-day B compset at the 24 time step where the ocean component is coupled 4 times. The CESM grid resolution was “ne30g16”, which corresponds to a 1° grid containing a total of 48602 horizontal grid points and 30 vertical levels for the atmosphere components, and 1° grid

https://doi.org/10.5194/gmd-2024-10
Preprint. Discussion started: 29 January 2024

containing 320×384 grid points and 60 vertical levels for the ocean components. Simulations were run with 16 CGs on the new Sunway. The default compiler on the new Sunway for the CESM1.3 is SW9 with -O2 optimization.

The GPU computing system we used for training the BGRU-AE model consists of Nvidia Tesla V100. Each Tesla V100 GPU contains 80 multithreaded streaming multiprocessors (SMs) and 16 GB of global DDR4 memory. Each SM contains 64 FP32 cores, 32 FP64 cores, and 8 Tensor cores (Kelly, 2010; Fuhrer et al., 2018).

As noted in 3.4, we use the processed training and validation datasets to train the BGRU-AE model. The reconstruction errors of the training and validation datasets are shown in **Fig. 6**. The values of optimal parameters in the BGRU-AE model are shown in **Table 3**.

We calculate the reconstruction errors after re-inputting the training datasets into the saved BGRU-AE model. The PDF of the reconstruction errors of the training datasets is shown in **Fig. 7**. Following **Fig. 7**, we take 0.05 as the threshold of reconstruction errors. Therefore, if the reconstruction error of one member of the testing datasets can be higher for those of maximum value of the training datasets, our tools for the consistency evaluation return the member as “failure”.

4.2 Non-climate changing modifications

To evaluate the consistency on heterogeneous many-core systems with mixed perturbations of software changes and hardware updates, the tool must accept the unavoidable perturbations caused by the heterogeneous hardware designs. The heterogeneous version of CESM on the new Sunway system is configured with the Zhang and McFarlane cumulus convection parameterization scheme (ZM scheme) (Zhang and Mcfarlane, 1995) of atmosphere components and K Profile parameterization scheme (KPP scheme; Large et al., 1994) of ocean components are in the heterogeneous computing. We input the simulation results of the heterogeneous version of CESM on the new Sunway system into the ESM-DCT, where other configurations are the same with training datasets such as -O2 compiling optimization option, as shown in **Table 4**. **Figure 8** shows the reconstruction errors of the testing datasets with the acceptable hardware-related perturbations. The result shows that the heterogeneous computing can not affect the consistency of the CESM simulation results and the tool can accept the perturbations caused by the heterogeneous designs.

Then, we focus on the influence of the acceptable compiling optimization option changes on the consistency of simulation results taking hardware-related perturbations into account. In the process of translating high-level programming language into machine language codes, different compiler optimization options can cause assembly code differences as different code execution order and/or different intermediate register floating-point precision, eventually causing nonidentical floating-point results. However, the modifications in -O0 and -O1 compiler optimization options can lead to nonidentical floating-point results but are not expected to be scientifically-changing. We expect that the testing datasets among -O0 and -O1 compiling optimization options can be consistent with the training datasets considering the ensemble distributions despite acceptable hardware-related perturbations involved. We input the simulation results of the heterogeneous version of CESM on the new

https://doi.org/10.5194/gmd-2024-10
Preprint. Discussion started: 29 January 2024


Geoscientific Model Development Discussions

Sunway system into the ESM-DCT with -O0 and -O1 compiling optimization options. The passing rate of the testing datasets among -O0 and -O1 on the new Sunway are shown in **Table 4**. **Figure 8** shows the reconstruction errors of the testing datasets with different compiling optimization options. The result shows that the tool can accept the mixed perturbations caused by the hardware designs and acceptable compiling optimization option changes.

4.3 Climate changing modifications

Our tool must successfully detect the inconsistency of the simulation results that are known to be climate-changing. Climate scientists provided a list of CAM input parameters thought to affect the climate in a non-trivial manner, which is used to detect changes to the simulation results that are known to be climate-changing in the CESM-ECT (Baker et al., 2015), such as *c0_lnd* and *c0_ocn*. In this study, we modify the values of input parameters in the ZM scheme of atmosphere components and then test whether or not our tool can detect the inconsistency caused by the model parameter changes using the ESM-DCT in the heterogeneous computing. Meanwhile, as noted in Section 3.3, the ESM-DCT must evaluate the consistency of the simulation results of the CESM with the unacceptable initial perturbations in the heterogeneous computing, where the ensembles are with the $O(10^{-6})$ perturbations of initial atmospheric temperature. We expect that the testing datasets with mixed perturbations caused by the climate changing modifications and hardware designs can not be consistent with the training datasets considering the ensemble distributions. We input the simulation results of the heterogeneous version of CESM on the new Sunway system into the ESM-DCT with model parameter and initial perturbations changes known to be climate-changing. The passing rates of the testing datasets with climate changing modifications on the new Sunway are shown in **Table 5**. The reconstruction errors of the testing datasets with climate changing modifications are shown in **Figure 9**. The results show that the tool can detect the climate changing modifications when taking hardware-related perturbations into account.

4.4 Unknown outcomes modifications

We present results for simulations in which we had less confidence in the expected outcome. For example, the effect of -O3 compiler optimization option was not known, because the CESM code base is large and level-three optimizations can be quite aggressive (Baker et al., 2015). We input the simulation results of the heterogeneous version of CESM on the new Sunway system into the ESM-DCT with -O3 compiler optimization option. The passing rate of the testing datasets with -O3 compiler optimization option on the new Sunway system is shown in **Table 6**. **Figure 10** shows the reconstruction errors of the testing datasets with -O3 compiler optimization option. The result shows that the effect of -O3 compiler optimization option in the new Sunway system is positive, which provides the references for the porting and optimization of CESM on the new Sunway system.

https://doi.org/10.5194/gmd-2024-10
Preprint. Discussion started: 29 January 2024


Geoscientific Model Development Discussions

ESMs need to use computational resources efficiently where mixed-precision approaches are emerging as a potential solution to help improve efficiency (Prims et al., 2019). ESMs in mixed precision programming must assess the simulation outputs to see if the results are accurate enough. In this study, we respectively set part of variables in the ZM parameterization scheme to the single precision, and then detect whether or not the mixed precision programming can affect the accuracy of the results using the ESCM-DCT. The passing rate of the testing datasets in the mixed precision programming on the new Sunway is shown in **Table 6**. **Figure 11** shows the reconstruction errors of the testing datasets in the mixed precision programming. The result shows that *rliq* and *pflx* variables can be set to single precision in CESM1.3 the heterogeneous version on the new Sunway. Our tool can serve as a rapid method for detecting correctness in the mixed precision programming to help ESMs benefit from a reduction of the precision of certain variables on the heterogeneous many-core HPC systems.

Then, our tool can detect sensitivity of input parameters, which is excluded to the input parameter list provided by the climate scientist thought to affect the climate in a non-trivial manner. In this study, we respectively modify the values of input parameters in the ZM scheme of atmosphere components and KPP scheme of ocean components and then test whether or not our tool can detect the inconsistency caused by the model parameter changes using the ESM-DCT, as shown in **Table 6**. **Figure 10** shows the reconstruction errors of the application datasets with the model parameter changes. The result shows that *ke*, *vdc_eq*, and *vdc_psim* variables are not sensitive to the configuration in this study, while the value of the *vdc1* variables should not be changed.

5 Summary and discussions

Numerical simulation advancements which demand tremendous computing power drive the progressive hardware upgrade of modern supercomputers. In terms of architecture, due to physical and heat limits, most of the supercomputing systems in the last decade came in the heterogeneous structure to improve the performance continuously, such as the new Sunway system. There exist differences in hardware designs between general-purpose and accelerator cores in heterogeneous many-core architecture computing environments, which causes the unavoidable computational perturbations and uncertainties whenever a accelerator core is involved. The computational perturbations caused by hardware designs and software or human errors can form a mixed perturbation computing environment, which affects the scientific consistency evaluation results for model verification. Hence, an efficient and objective scientific consistency test approach on the heterogeneous many-core architectures is urgently demanded, which can accept the influence of the computational perturbations for further detection of software or human errors generated in optimizing and developing the ESMs.

In this study, we develop a deep learning-based consistency test approach for the software verification of CESM on the new Sunway system, referred to as ESM-DCT. First, we generate a series of ensembles of short-term simulations on the

https://doi.org/10.5194/gmd-2024-10
Preprint. Discussion started: 29 January 2024

multi-component as the datasets which capture the natural variability in the modeled climate system. Then, we train the BGRU-AE model to study the variable features from the ensembles of atmosphere and ocean components, which makes unavoidable computational perturbations caused by hardware designs accepted for further detection of software or human errors. Finally, we use the software tool to evaluate whether or not a new CESM configuration facing with the mixed perturbation composed of hardware designs and software or human errors (e.g. compiling optimization option changes, mixed precision programming, and model parameter changes) in the heterogeneous computing is consistent with the original "trusted" configuration in the homogeneous computing. Our current efforts increase the confidence in detecting and reducing errors in the development and optimization of ESMs on the heterogeneous many-core systems.

Although the focus of this study is on the specific new Sunway system, the deep learning-based consistency test approach provides a protocol for other heterogeneous many-core systems such as GPU-based high-performance computing systems, and future heterogeneous HPC systems with hardware updated for more affordable energy consumption. Besides ESMs, the heterogeneous HPC systems can be applied to many other scientific research fields such as biotechnology and new materials, etc., where the scientific computing results also suffer from the uncertainties due to computational perturbations caused by the heterogeneous architecture. The deep learning-based consistency test approach provides an efficient and objective approach for verifying the reliability of the development and optimization of scientific computing models on the heterogeneous HPC systems. Furthermore, our future endeavors involve an enhancement of spatial pattern evaluation, extending beyond global spatial means to achieve more nuanced feature extraction. Subsequent research will focus on refining convolutional neural networks and attention mechanisms within the deep-learning model architecture, along with an augmentation of model parameters to mitigate overfitting concerns.

Code and data availability

Codes, data and scripts used to run the models and produce the figures in this work are available on the Zenodo site (https://doi.org/10.5281/zenodo.10467529, Yu et al., 2024) or by sending a written request to the corresponding author (Shaoqing Zhang, szhang@ouc.edu.cn).

Author contributions

Yangyang Yu is responsible for all plots, initial analysis and some writing; Shaoqing Zhang leads the project, organizes and refines the paper; Haohuan Fu and Dexun Chen provide significant discussions and inputs for the whole research; all other co-authors make equal contributions by wording discussions, comments and reading proof.

https://doi.org/10.5194/gmd-2024-10
Preprint. Discussion started: 29 January 2024

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgments

This research is supported by Science and Technology Innovation Project of Laoshan Laboratory (grant no. LSKJ202202202), the National Key R&D Program of China (grant no. 2022YFE0106400), the National Natural Science Foundation of China (grant no. 41830964), Shandong Province's "Taishan" Scientist Program (grant no. ts201712017), Postdoctoral Fellowship Program of CPSF (grant no. GZC20232491). All numerical experiments are performed on the homogeneous and heterogeneous supercomputing platforms at Laoshan Laboratory.

References

Asselin, R.: Frequency filter for time integrations. Mon Weather Rev, 100, 487-490, https://doi.org/10.1175/1520-0493(1972)100%3c0487:FFFTI%3e2.3.CO;2, 1972.

Arteaga, A., Fuhrer, O., and Hoefler, T.: Designing Bit-Reproducible Portable High-Performance Applications, 2014 IEEE International Parallel & Distributed Processing Symposium (IPDPS), 1235-1244, https://doi.org/10.1109/IPDPS.2014.127, 2014.

Baker, A. H., Hammerling, D. M., Levy, M. N., Xu, H., Dennis, J. M., Eaton, B. E., Edwards, J., Hannay, C., Mickelson, S. A., Neale, R. B., Nychka, D., Shollenberger, J., Tribbia, J., Vertenstein, M., and Williamson, D.: A new ensemble-based consistency test for the Community Earth System Model (pyCECT v1.0), Geosci Model Dev, 8, 3823-3859, https://doi.org/10.5194/gmd-8-2829-2015, 2015.

Baker, A. H., Hu, Y., Hammerling, D. M., Tseng, Y.-H., Xu, H., Huang, X., Bryan, F. O., and Yang, G.: Evaluating statistical consistency in the ocean model component of the Community Earth System Model (pyCECT v2.0), Geosci Model Dev, 9, 2391-2406, https://doi.org/10.5194/gmd-9-2391-2016, 2016.

Carson II, J. S.: Model verification and validation, Proceedings of the 2002 Winter Simulation Conference, San Diego, USA, 52-58, https://doi.org/10.1109/WSC.2002.1172868, 2002.

Cho, K., Merrienboer, B. V., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y.: Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, Comp Sci, 1-15, https://doi.org/10.48550/arXiv.1406.1078, 2014.

Clune, T. and Rood, R.: Software testing and verification in climate model development, IEEE Softw, 28, 49-55, https://doi.org/10.1109/MS.2011.117, 2011.

https://doi.org/10.5194/gmd-2024-10
Preprint. Discussion started: 29 January 2024

Düben, P. D., Joven, J., Lingamneni, A., McNamara, H., Micheli, G. D., Palem, K. V., Palmer, T.N.: On the use of inexact, pruned hardware in atmospheric modelling, Phil Trans R Soc, A 372: 20130276, http://dx.doi.org/10.1098/rsta.2013.0276, 2014.

Düben, P. D., Subramanian, A., Dawson, A., and Palmer T. N.: A study of reduced numerical precision to make superparametrisation more competitive using a hardware emulator in the OpenIFS model, J Adv Model Earth Syst, 9, 566-584, https://doi.org/10.1002/2016MS000862, 2017.

Easterbrook, S. M. and Johns, T. C.: Engineering the software for understanding climate change, Comput Sci Eng, 11, 65-74, https://doi.org/10.1109/MCSE.2009.193, 2009.

Flato, G. M.: Earth system models: an overview, WIREs Clim Change, 2, 783-800, https://doi.org/10.1002/wcc.148, 2011.

Fotiadou, K., Velivassaki, T. H., Voulkidis, A., Skias, D., Tsekeridou, S., and Zahariadis, T.: Network Traffic Anomaly Detection via Deep Learning, Information, 12, 215. https://doi.org/10.3390/info12050215, 2021.

Fu, H., Liao, J., Yang, J., Wang, L., Song, Z., Huang, X., Yang, C., Xue, W., Liu, F., Qiao, F., Zhao, W., Yin, X., Hou, C., Zhang, C., Ge, W., Zhang, J., Wang, Y., Zhou, C., and Yang, G.: The sunway taihulight supercomputer: system and applications, Sci China Inf Sci, 59, 072001, https://doi.org/10.1007/s11432-016-5588-7, 2016.

Fu, H., Liao, J., Ding, N., Duan, X., Gan, L., Liang, Y., Wang, X., Yang, J., Zheng, Y., Liu, W., Wang, L., and Yang, G.: Redesigning cam-se for peta-scale climate modeling performance and ultra-high resolution on sunway taihulight, International conference for high performance computing, networking, storage and analysis, 2017a.

Fu, H., Liao, J., Xue, W., Wang, L., Chen, D., Gu, L., Xu, J., Ding, N., Wang, X., He, C., Xu, S., Liang, Y., Fang, J., Xu, Y., Zheng, W., Xu, J., Zheng, Z., Wei, W., Ji, X., Zhang, H., Chen, B., Li, K., Huang, X., Chen, W., and Yang, G.: Refactoring and optimizing the community atmosphere model (CAM) on the sunway taihu-light supercomputer. International Conference for High Performance Computing, Networking, Storage and Analysis, 2017b.

Fuhrer, O., Chadha, T., Hoefler, T., Kwasniewski, G., Lapillonne, X., Leutwyler, D., Lüthi, D., Osuna, C., Schär, C., Schulthess, T. C., and Vogt, H.: Near-global climate simulation at 1 km resolution: establishing a performance baseline on 4888 GPUs with COSMO 5.0, Geosci Model Dev, 11, 1665-1681, https://doi.org/10.5194/gmd-11-1665-2018, 2018.

Gnanadesikan, A.: A simple predictive model for the structure of the oceanic pycnocline, Science, 283, 2077-2079, https://doi.org/10.1126/science.283.5410.2077, 1999.

Gong, D., Liu, L., Le, V., Saha, B., Mansour, M. R., Venkatesh, S. and Hengel, A. V.-D.: Memorizing Normality to Detect Anomaly: Memory-augmented Deep Autoencoder for Unsupervised Anomaly Detection, 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Korea (South), https://doi.org/10.1109/ICCV.2019.00179, 2019.

Gu, J., Feng, J., Hao, X., Fang, T., Zhao, C., An, H., Chen, J., Xu, M., Li, J., Han, W., Yang, C., Li, F., and Chen, D.: Establishing a non-hydrostatic global atmospheric modeling system at 3-km horizontal resolution with aerosol

https://doi.org/10.5194/gmd-2024-10
Preprint. Discussion started: 29 January 2024

feedbacks on the Sunway supercomputer of China. Sci Bull, 67, 1170-1181, https://doi.org/10.1016/j.scib.2022.03.009, 2022.

Hinton, G. E. and Salakhutdinov, R. R.: Reducing the Dimensionality of Data with Neural Networks, Science, 313, 504-507, https://doi.org/10.1126/science.1127647, 2006.

Hochreiter, S. and Schmidhuber, J.: Long short-term memory. Neural Comput, 9, 1735-1780, https://doi.org/10.1162/neco.1997.9.8.1735, 1997.

Hurrell, J. W., Holland, M. M., Gent, P. R., Ghan, S., Kay, J. E., Kushner, P. J., Lamarque, J. F., Large, W. G., Lawrence, D., Lindsay, K., Lipscomb, W. H., Long, M. C., Mahowald, N., Marsh, D. R., Neale, R. B., Rasch, P., Vavrus, S., Vertenstein, M., Bader, D., Collins, W. D., Hack, J. J., Kiehl, J., and Marshall, S.: The Community Earth System Model: a framework for collaborative research, B Am Metereol Soc, 94, 1339-1360, https://doi.org/10.1175/BAMS-D-12-00121.1, 2013.

Kelly, R. C.: GPU Computing for Atmospheric Modeling, Comput Sci Eng, 12, 26-33, https://doi.org/10.1109/MCSE.2010.26, 2010.

Kish, L. B.: End of Moore's law: thermal (noise) death of integration in micro and nano electronics, Phys Lett A, 305, 144-149, https://doi.org/10.1016/S0375-9601(02)01365-8, 2002.

Large, W. G., Mcwilliams J. C., and Doney S. C.: Oceanic vertical mixing: A review and a model with a nonlocal boundary layer parameterization, Rev Geophys, 32, 363-403, https://doi.org/10.1029/94RG01872, 1994.

Liao, X., Xiao, L., Yang, C., and Lu, Y.: Milkyway-2 supercomputer: system and application, Front Comput Sci, 8, 345-356, https://doi.org/10.1007/s11704-014-3501-3, 2014.

Liu, X., Kruger, U., Littler, T., Xie, L., and Wang, S.: Moving window kernel pca for adaptive monitoring of nonlinear processes, Chemom Intell Lab Syst, 96: 132-143, https://doi.org/10.1016/j.chemolab.2009.01.002, 2009.

Lorenz, E. N. Deterministic non-periodic flow. J Atmos Sci, 20, 130-141, https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2, 1963.

Lukosevicius, M., Jaeger, H.: Reservoir computing approaches to recurrent neural network training. Comput Sci Rev, 3, 127-149, https://doi.org/10.1016/j.cosrev.2009.03.005, 2009.

Maya, S., Ueno, K., and Nishikawa, T.: dLSTM: a new approach for anomaly detection using deep learning with delayed prediction, Int J Data Sci Anal, 8, 137-164, https://doi.org/10.1007/s41060-019-00186-0, 2019.

Milroy, D. J., Baker, A. H., Hammerling, D. M., Dennis, J. M., Mickelson, S. A., and Jessup, E. R.: Towards Characterizing the Variability of Statistically Consistent Community Earth System Model Simulations, Procedia Computer Science, 80, 1589-1600, https://doi.org/10.1016/j.procs.2016.05.489, 2016.

https://doi.org/10.5194/gmd-2024-10
Preprint. Discussion started: 29 January 2024

Milroy, D. J., Baker, A. H., Hammerling, D. M., and Jessup, E. R.: Nine time steps: ultra-fast statistical consistency testing of the Community Earth System Model (pyCECT v3.0), Geosci Model Dev, 11, 697-711, https://doi.org/10.5194/gmd-11-697-2018, 2018.

Prims, O. T., Acosta, M. C., Moore, A. M., Castrillo, M., Serradell, K., Cortés, A., and Doblas-Reyes, F. J.: How to use mixed precision in ocean models: exploring a potential reduction of numerical precision in NEMO 4.0 and ROMS 3.6, Geosci Model Dev, 12, 3135-3148, https://doi.org/10.5194/gmd-12-3135-2019, 2019.

Rosinski, J. M., and Williamson, D. L.: The accumulation of rounding errors and port validation for global atmospheric models, SIAM J Sci Comput, 18, 552-564, https://doi.org/10.1137/S1064827594275534, 1997.

Sansom, P. G., Stephenson, D. B., Ferro, C. A. T., Zappa, G., and Shaffery, L.: Simple uncertainty frameworks for selecting weighting schemes and interpreting multimodel ensemble climate change experiments, J Clim, 26, 4017-4037, https://doi.org/10.1175/JCLI-D-12-00462.1, 2013.

Song, T., Han, N., Zhu, Y., Li, Z., Li, Y., Li, S., and Peng, S.: Application of deep learning technique to the sea surface height prediction in the South China Sea, Acta Oceanol Sin, 40, 68-76. https://doi.org/10.1007/s13131-021-1735-0, 2021.

Su, Y., Kuo, C.-C. J.: On extended long short-term memory and dependent bidirectional recurrent neural network, Neurocomputing, 356, 151-161, https://doi.org/10.1016/j.neucom.2019.04.044, 2019.

Sun, D., Wulff, J., Sudderth, E. B., Pfister, H., and Black, M. J.: A Fully-Connected Layered Model of Foreground and Background Flow, 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, https://doi.org/10.1109/CVPR.2013.317, 2013.

Vazhkudai, S. S., de Supinski, B. R., Bland, A. S., Geist, A., Sexton, J., Kahle, J., Zimmer, C. J., Atchley, S., Oral, S., Maxwell, D. E., Vergara Larrea, V. G., Bertsch, A., Goldstone, R., Joubert, W., Chambreau, C., Appelhans, D., Blackmore, R., Casses, B., Chochia, G., Davison, G., Ezell, M. A., Gooding, T., Gonsiorowski, E., Grinberg, L., Hanson, B., Hartner, B., Karlin, I., Leininger, M. L., Leverman, D., Marroquin, C., Moody, A., Ohmacht, M., Pankajakshan, R., Pizzano, F., Rogers, J. H., Rosenburg, B., Schmidt, D., Shankar, M., Wang, F., Watson, P., Walkup, B., Weems, L. D., and Yin, J.: The design, deployment, and evaluation of the coral pre-exascale systems, International conference for high performance computing, networking, storage, and analysis, Dallas, USA, 2018.

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P. A.: Extracting and Composing Robust Features with Denoising Autoencoders, Proceedings of the 25th international conference on Machine learning, Finland, https://doi.org/10.1145/1390156.1390294, 2008.

Yu, W., Kim, Y., and Mechefske, C.: Remaining useful life estimation using a bidirectional recurrent neural network based autoencoder scheme, Mech Syst Signal Pr, 129, 764-780, https://doi.org/10.1016/j.ymssp.2019.05.005, 2019.

https://doi.org/10.5194/gmd-2024-10
Preprint. Discussion started: 29 January 2024

Yu, Y., Zhang, S., Fu, H., Wu, L., Chen, D., Gao, Y., Wei, Z., Jia, D., and Lin, X.: Characterizing uncertainties of Earth system modeling with heterogeneous many-core architecture computing, Geosci Model Dev, 15, 6695-6708, https://doi.org/10.5194/gmd-15-6695-2022, 2022.

Zhang, G. J., and Mcfarlane, N. A.: Sensitivity of climate simulations to the parameterization of cumulus convection in the Canadian climate centre general circulation model, Atmos ocean, 33, 407-446, https://doi.org/10.1080/07055900.1995.9649539, 1995.

Zhao, R., Wang, D., Yan, R., Mao, K., Shen, F., and Wang, J.: Machine Health Monitoring Using Local Feature-based Gated Recurrent Unit Networks, IEEE Transactions on Industrial Electronics, 65, 1539-1548, https://doi.org/10.1109/TIE.2017.2733438, 2017.

Zhang, S.: Impact of observation-optimized model parameters on decadal predictions: simulation with a simple pycnocline prediction model, Geophys Res Lett, 38, L02702. https://doi.org/10.1029/2010GL046133,2011.

Zhang, S., Fu, H., Wu, L., Li, Y., Wang, H., Zeng, Y., Duan, X., Wan, W., Wang, L., Zhuang, Y., Meng, H., Xu, K., Xu, P., Gan, L., Liu, Z., Wu, S., Chen, Y., Yu, H., Shi, S., Wang, L., Xu, S., Xue, W., Liu, W., Guo, Q., Zhang, J., Zhu, G., Tu, Y., Edwards, J., Baker, A. H., Yong, J., Yuan, M., Yu, Y., Zhang, Q., Liu, Z., Li, M., Jia, D., Yang, G., Wei, Z., Pan, J., Chang, P., Danabasoglu, G., Yeager, S., Rosenbloom, N., and Guo, Y.: Optimizing High-Resolution Community Earth System Model on a Heterogeneous Many-Core Supercomputing Platform (CESMHR_sw1.0), Geosci Model Dev, 13, 4809-4829, https://doi.org/10.5194/gmd-2020-18, 2020.

Zhang, S., Xu, S., Fu, H., Wu, L., Liu, Z., Gao, Y., Zhao, C., Wan, W., Wan, L., Lu, H., Li, C., Liu, Y., Lv, X., Xie, J., Yu, Y., Gu, J., Wang, X., Zhang, Y., Ning, C., Fei, Y., Guo, X., Wang, Z., Wang, X., Wang, Z., Qu B., Li, M., Zhao, H., Jiang, Y., Yang, G., Lu, L., Wang, H., An, H., Zhang, X., Zhang, Y., Ma, W., Yu, F., Xu, J., Lin, X., and Shen, X.: Toward Earth system modeling with resolved clouds and ocean submesoscales on heterogeneous many-core HPCs, Natl Sci Rev, 10, nwad069, https://doi.org/10.1093/nsr/nwad069, 2023.

Zhao, Y., Deng, X., Zhang, S., Liu, Z., and Liu, C.: Sensitivity determined simultaneous estimation of multiple parameters in coupled models: part I-based on single model component sensitivities, Clim Dynam, 53, 5349-5373, https://doi.org/10.1007/s00382-019-04865-3, 2019.

https://doi.org/10.5194/gmd-2024-10
Preprint. Discussion started: 29 January 2024

Geoscientific Model Development
Discussions
Open Access
EGU

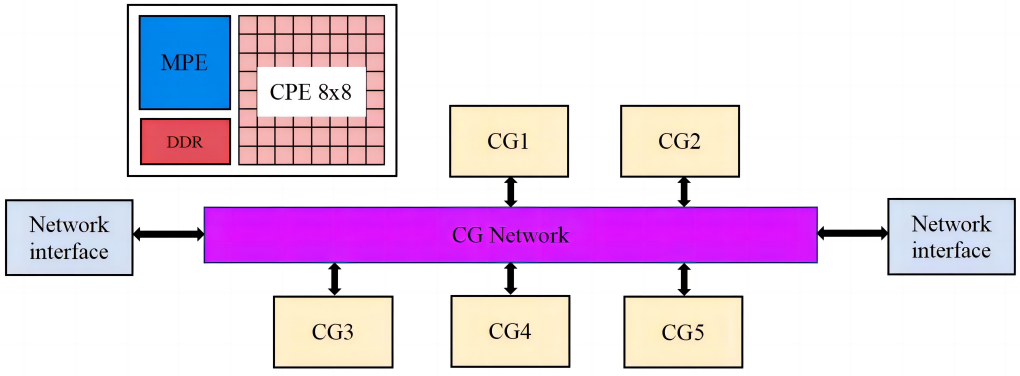


Figure 1: The illustration of the general architecture of the Sunway SW26010P processor. Each processor consists of 6 Core Groups, and each Core Group includes DDR memory, a general-purpose core (MPE), and 64 accelerator cores (CPEs). 6 Core Groups (CGs) are linked together by the CG Network, and the whole many-core processor is linked with other processors by the network interface (Courtesy to Gu et al., 2022).

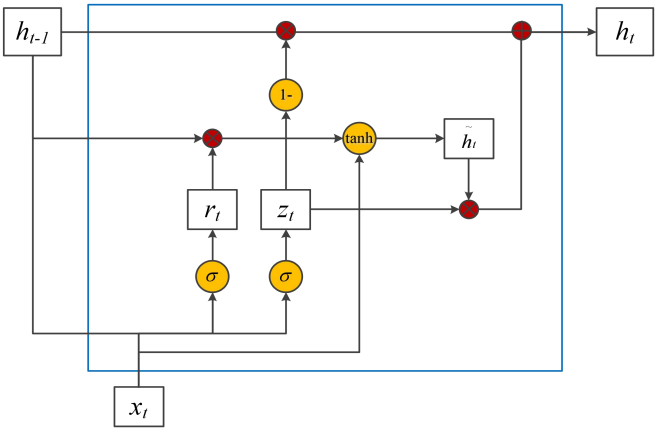
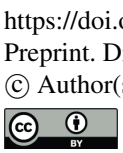


Figure2: The illustrations of GRU models.

https://doi.org/10.5194/gmd-2024-10
Preprint. Discussion started: 29 January 2024



Geoscientific Model Development
Discussions
Open Access
EGU

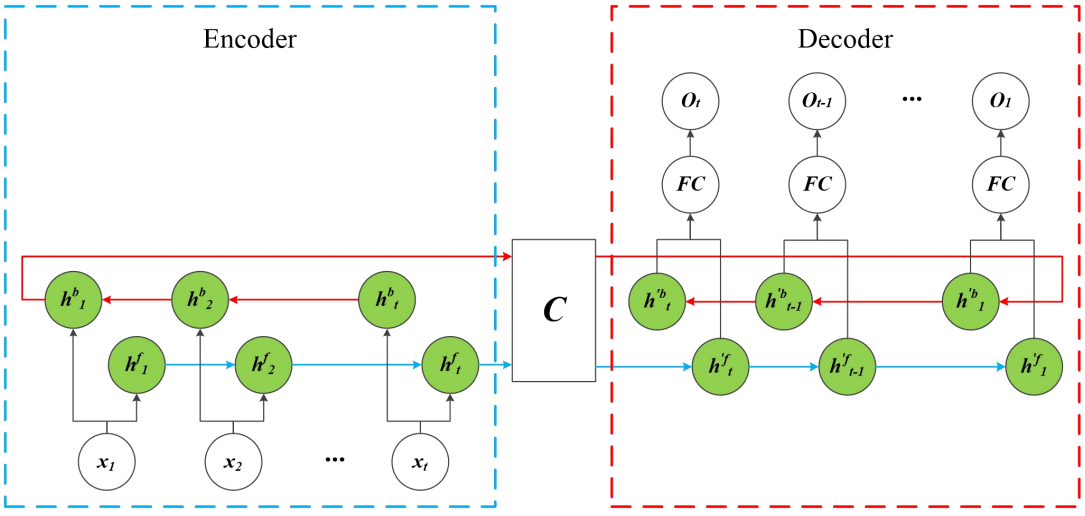


Figure 3: The illustrations of BGRU-AE models. The blue dashed box represents the encoder, and the red dashed box represents the decoder. The short code *C* is the output of the encoder BGRU. The final result *O* is the output of the decoder BGRU performing FC.

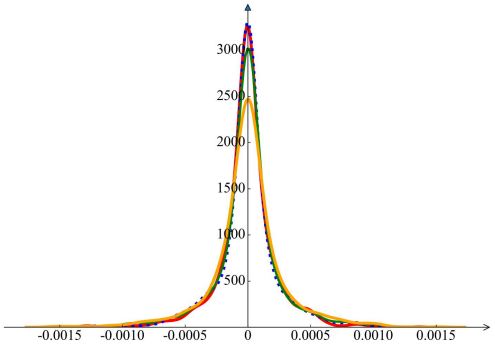


Figure 4: The PDFs of atmosphere temperature with different orders of magnitude of perturbations. The PDFs with the O(10^{-14}), O(10^{-7}), O(10^{-6}), and O(10^{-5}) perturbations are represented by the red, blue-dotted, green, and orange lines.

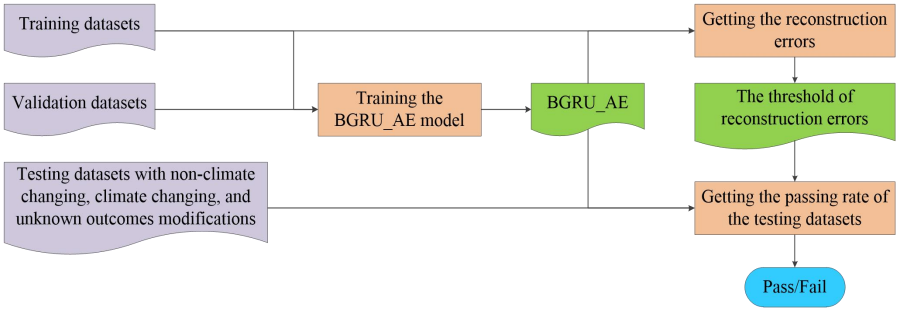


Figure 5: The flow chart of the ESM-DCT tool for the consistency evaluation on the heterogeneous many-core HPC systems.

https://doi.org/10.5194/gmd-2024-10
Preprint. Discussion started: 29 January 2024

Geoscientific Model Development
Discussions
Open Access
EGU

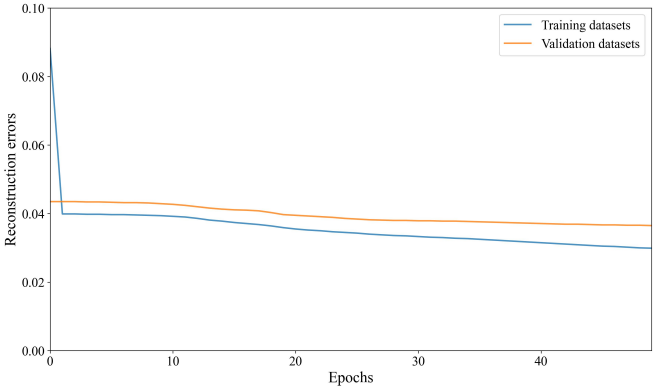


Figure 6: The reconstruction errors of the training datasets and validation datasets.

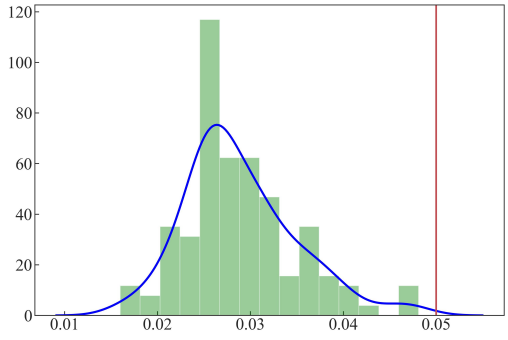


Figure 7: The PDF of the reconstruction errors of the training datasets. The red line is the threshold of reconstruction errors.

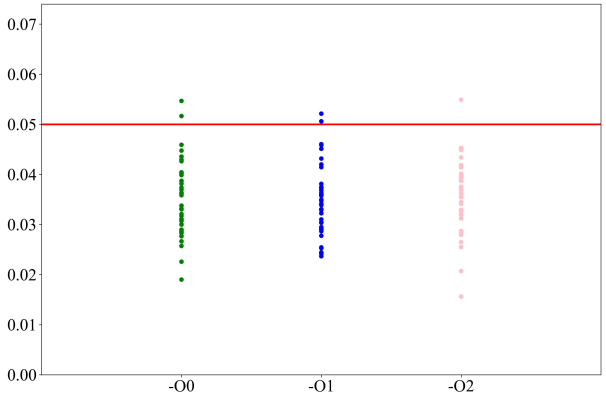


Figure 8: The reconstruction errors of the datasets with different compiling optimization options. The red line is the threshold of reconstruction errors of the training datasets. -O2 is the testing datasets with the acceptable hardware-related perturbations. -O0 and -O1 are the testing datasets with mixed perturbations caused by the hardware designs and compiling optimization option changes.

https://doi.org/10.5194/gmd-2024-10
Preprint. Discussion started: 29 January 2024

Geoscientific Model Development Discussions

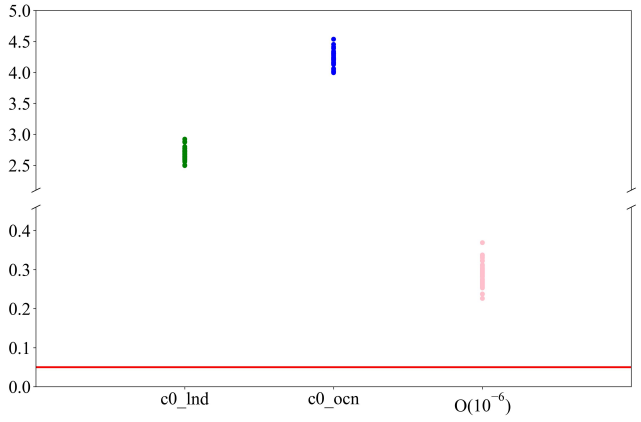


Figure 9: The reconstruction errors of the datasets with the climate changing modifications. The red line is the threshold of reconstruction errors of the training datasets. c0_lnd and c0_ocn are the testing datasets with mixed perturbations caused by the hardware designs and model parameter changes known to be climate-changing. O(10^{-6}) is the testing datasets with mixed perturbations caused by the hardware designs and initial perturbations changes known to be climate-changing.

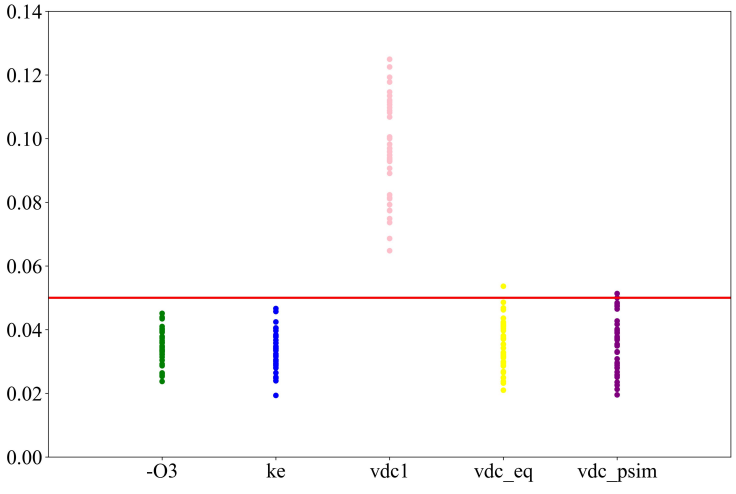


Figure 10: The reconstruction errors of the datasets with -O3 compiling optimization option and the model parameter changes. The red line is the threshold of reconstruction errors of the training datasets. -O3 is the testing datasets with mixed perturbations caused by the hardware designs and compiling optimization option changes. ke, vdc1, vdc_eq, and vdc_psim are the testing datasets with mixed perturbations caused by the hardware designs and model parameter changes.

https://doi.org/10.5194/gmd-2024-10
Preprint. Discussion started: 29 January 2024

Geoscientific Model Development Discussions

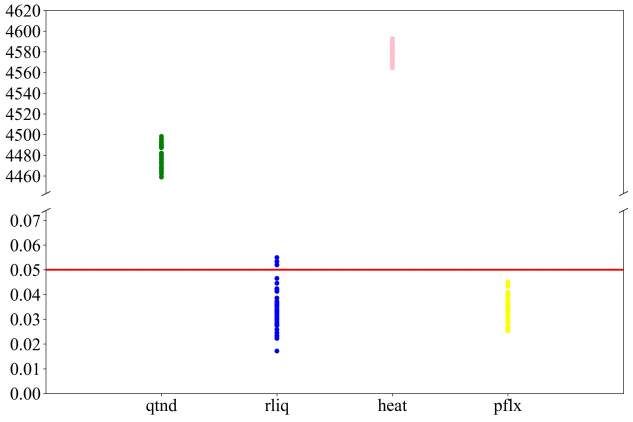


Figure 11: The reconstruction errors of the datasets in the mixed precision programming. The red line is the threshold of reconstruction errors of the training datasets. qtnd, rliq, heat, pflx are the testing datasets with mixed perturbations caused by the hardware designs and variable precision changes.

575

Table 1: Results of the ECT and DCT for the 5VCCM with GNU/Intel compilers and heterogeneous computing

Test name	Number of PCs failing at least two runs	ECT results	Passing rate of DCT	DCT results
GNU	1	Pass	97.5%	Pass
Intel	2	Failure	90%	Pass
Heterogeneous computing	2	Failure	100%	Pass

Table 2: The list of datasets for the ESM-DCT

Datasets	Test name	Functions	Descriptions
Training datasets	Training the BGRU-AE model	Training the model parameters	Simulations in the homogeneous computing
Validation datasets			Simulations in the homogeneous computing
Testing datasets	Non-climate changing modifications	Testing the acceptable hardware-related perturbations	Simulations in the heterogeneous computing
	Non-climate changing modifications	Testing the acceptable compiling optimization option adjustments	Simulations in the heterogeneous computing with -O0 and -O1 compiling optimization options
	Climate changing modifications	Testing the unacceptable initial perturbations	Simulations in the heterogeneous computing with $O(10^{-6})$ initial atmosphere perturbations
	Climate changing modifications	Testing the unacceptable CESM model parameter adjustments	Simulations in the heterogeneous computing with unacceptable CESM model parameter adjustments
	Unknown outcomes modifications	Testing the compiling optimization option adjustments with unknown effects	Simulations in the heterogeneous computing with -O3 compiling optimization option
	Unknown outcomes modifications	Testing the mixed precision programming	Simulations in the heterogeneous computing in the mixed precision programming
	Unknown outcomes modifications	Testing the CESM model parameter adjustments with unknown effects	Simulations in the heterogeneous computing with CESM model parameter adjustments

https://doi.org/10.5194/gmd-2024-10
Preprint. Discussion started: 29 January 2024

Geoscientific Model Development Discussions

Table 3: The value of optimal parameters in the BGRU-AE model.

Item	Value
Learning rate	0.001
Epochs	50
Layer number	1
Hidden-size	16
Batch-size	1
Dropout	0

Table 4: The results of the datasets with different compiling optimization options in the ESM-DCT.

Test name	Passing rate	ESM-DCT results
-O0	95%	Pass
-O1	95%	Pass
-O2	97.5%	Pass

Table 5: The results of the testing datasets with the climate changing modifications in the ESM-DCT.

Test name	Descriptions	Original value	Modified value	Subroutines	Passing rate	ESM-DCT results
c0_lnd	Autoconversion coefficient over land	0.0059	0.0039	ZM scheme	0%	Pass
c0_ocn	Autoconversion coefficient over ocean	0.045	0.025	ZM scheme	0%	Pass
$O(10^{-6})$	Initial atmosphere perturbations	-	-	-	0%	Pass

https://doi.org/10.5194/gmd-2024-10
Preprint. Discussion started: 29 January 2024

Geoscientific Model Development Discussions

Table 6: The results of the testing datasets with unknown outcomes in the ESM-DCT.

Test name	Descriptions	Original value	Modified value	Subroutines	Passing rate	ESM-DCT results
Changes in compiler optimization option						
-O3	Compiling optimization option	-	-	ZM scheme	100%	Pass
Changes in variable precision						
qtnd	Specific humidity tendency	Double precision	Single precision	ZM scheme	0%	Failure
rliq	Reserved liquid for energy integrals	Double precision	Single precision	ZM scheme	92.5%	Pass
heat	Dry static energy tendency	Double precision	Single precision	ZM scheme	0%	Failure
pflx	Scattered precip flux	Double precision	Single precision	ZM scheme	100%	Pass
Changes in model parameter						
ke	Tunable evaporation efficiency	1.0E-6	2.0E-6	ZM scheme	92.5%	Pass
vdc1	Background diffusivity	0.16	0.26	KPP scheme	0%	Failure
vdc_eq	Equatorial diffusivity	0.01	0.02	KPP scheme	97.5%	Pass
vdc_psim	Maximum PSI-induced diffusivity	0.13	0.15	KPP scheme	97.5%	Pass