

ARTICLE



<https://doi.org/10.1038/s43247-020-00085-4>

OPEN

Limits of reproducibility and hydrodynamic noise in atmospheric regional modelling

Beate Geyer ¹, Thomas Ludwig^{2,3} & Hans von Storch ^{1,3}✉

Reproducibility of research results is a fundamental quality criterion in science; thus, computer architecture effects on simulation results must be determined. Here, we investigate whether an ensemble of runs of a regional climate model with the same code on different computer platforms generates the same sequences of similar and dissimilar weather streams when noise is seeded using different initial states of the atmosphere. Both ensembles were produced using a regional climate model named COSMO-CLM5.0 model with ERA-Interim forcing. Divergent phase timing was dependent on the dynamic state of the atmosphere and was not affected by noise seeded by changing computers or initial model state variations. Bitwise reproducibility of numerical results is possible with such models only if everything is fixed (i.e., computer, compiler, chosen options, boundary values, and initial conditions) and the order of mathematical operations is unchanged between program runs; otherwise, at best, statistically identical simulation results can be expected.

¹ Helmholtz-Zentrum Geesthacht, Centre for Materials and Coastal Research, Geesthacht, Germany. ² Deutsches Klimarechenzentrum, Hamburg, Germany.

³ Universität Hamburg, Hamburg, Germany. ✉email: hvonstorch@web.de

Models are executed as software packages on computers, which are considered deterministic systems. Notably, if you rerun a program with identical inputs, it will give an identical result, which represents a considerable benefit during model development, which is essentially software development because errors in the code can be detected easily; specifically, the exact reproduction of a program run enables us to track the source of a program error. In this sense, computer experiments are more precise than real-world experiments, where a single experiment can never be exactly reproduced. However, the deterministic quality of a computer is difficult to guarantee because it requires a special style of programming and is potentially disrupted as soon as you change one component of the system or the system itself¹. Consequently, the results of numerical simulations may differ.

Hydrodynamical systems such as the atmosphere and the ocean exhibit unprovoked variability in models if the spatial resolution allows for turbulent dynamics. Two simulations that are identical in setup but with slightly different initial values or other physically insignificant differences can quickly lead to statistically identical results but different trajectories. This finding has been observed in global atmospheric general circulation models since the 1970s². When the resolution of global ocean models is at the scale of eddy formation, the same finding can occur for ocean models (e.g.³). In the case of limited-area models, the issue was long disregarded because it was falsely assumed that the boundary values would suppress such divergence; however, this is not the case for regional dynamical models of the atmosphere⁴ or the ocean⁵. In this case, models can display divergence in phase space in certain periods, while in other time periods, all trajectories may remain close to each other.

We refer to such unprovoked differences as “noise”. This term is ambiguous, as it is commonly used in different fields of science and spans very different concepts in different disciplines. Originally from acoustics and electrical engineering, the term “noise” has found its way into many disciplines, including climate science and computer sciences.

- In climate science, the term refers to unprovoked variability – that is, variations that cannot be traced back to variable external factors. The famous Lorenz system, which considers the butterfly effect, only requires a miniscule disturbance to lead the system into very different states. When a regional atmospheric or oceanic model is initiated with observed states that are a day or more apart, the trajectory of the system will show at a later and potentially much later time, periods of divergence (“intermittent divergence in the phase space”⁶). In this case, the stochasticity is rooted in the very high dimension of the problem and the numerous nonlinear terms. The presence of internal variability is a property of the system.
- In computer science, numbers are stored with a finite number of digits. When using these numbers in codes, rounding errors occur and lead to small deviations from a mathematically accurate solution. This is independent of the chosen number of bits for the representation of numbers. With more bits and thus higher numerical precision, the differences occur later and have a smaller effect on the simulated climate. Note that more bits require more hardware resources and more energy for performing the simulation. Thus, the actual chosen precision is always a compromise between potential error and costs. Additionally, when the sequence of operations is modified, which is often the case with code or operation optimizations for parallel computing systems, or any minor other changes are implemented, different numbers will be generated (see, e.g.⁷).

Our hypothesis is that two ensembles generated by the same model that have the same boundary values for the same time and region but have different initial values or are run on different platforms but otherwise kept unmodified will be statistically identical. We expect that the episodes of divergent trajectories will occur at the same time in both cases and will have the same intensity – irrespective of the characteristics of the noise seeding process. We address this issue by comparing the performance of the same regional atmospheric model with two approaches: one that uses slightly shifted initial states and another that uses different computing platforms. Both methods lead to “noise”, and we will see that, through dynamic processes, this noise will lead to noteworthy variability in the results. It is important to note that noise induced by changing the computer platform cannot be quantified or controlled. Such noise is the effect of small variations in the execution order of mathematical operations in the processor hardware and the basic software libraries, which are opaque to the model developer. Precision issues of internal number representations will potentially have an impact but are beyond the control of the programmer.

The presence of unprovoked variability is significant, as it makes numerical experiments with models subject to stochastic variations associated with separating any signal of interest from the unavoidable internal variability, i.e., noise. Before claiming that an intervention, such as a change in boundary values or the parameterization scheme, has an effect, an ensemble of simulations must be performed, followed by an analysis of the signal-to-noise ratio of the impact on the simulation results^{2,4,8,9}. The signal to noise problem has been a well-known problem in global systems since the 1970s and in regional since the 1990s. A variety of statistical methods for identifying this issue was introduced into meteorology many years ago (c.f.¹⁰).

In inter-institutional cooperation the same code is occasionally executed on different computer platforms. Additionally, identical codes may be run at the same institutions on different platforms, such as when a new system is installed and the previously used platform is decommissioned. When performing numerical simulations and experiments on different platforms, the results of a simulation should not be affected by such a change. This condition is usually considered synonymous with having bit-identical results. However, as previously mentioned, this condition can only be guaranteed on a given parallel computer with a fixed software stack (including a selected fixed precision for all variables). Changing one of these components will potentially lead to different numbers. This approach is not necessarily identical to obtaining a different scientific result. However, if numbers vary, the results must be validated by different means. This task is identical to comparing the results of various real-world experiments.

How can we obtain the maximum bitwise reproducibility in computer experiments? Since different platforms, or different compilers, may also have miniscule differences, for instance, executing sums in different sequences¹¹, a change in the computer platform will also introduce small variations that can be physically insignificant but ubiquitous. Note that the technical details are unknown to the programmer and thus cannot be compensated for. Based on the “stochastic climate model”¹², we suggest that the hydrodynamics of the system will generate an ensemble of different trajectories, as in the case of a change in the initial state.

We emphasize the bi-disciplinary aspect of our work. The fact that miniscule changes in the code or in the process of executing a code lead to differences in the outcome is well known (e.g.^{13,14}); moreover, the fact that limited area models for atmospheric dynamics can generate macroscopic variations as a response to microscopic disturbances is equally known (e.g.^{3,6,11}). The

significance of this paper is the demonstration that disturbances do not depend on the initial disturbance or on the change of the employed computer; rather, any presence of a “noise seed” is sufficient to generate similar variations in terms of time and intensity over the course of the simulation independent of the time since initialization. We show that certain unstable atmospheric states cause the model to create noteworthy but intermittent variations. These variations do not represent “errors” but possible and consistent states, and the simulated time development must be considered a random variable.

In this paper, we consider seeded noise by minuscule changes in the initials states and by employing the same code on different platforms and are not considering changes of parameter settings, parametrizations or number precision, which may lead to systematically different simulations. To test deviations caused by such changes in the source code from an existing ensemble, one could use appropriate test tools, e.g., pyCECT¹⁵.

Results

Simulations with the regional climate model COSMO-CLM (COSMO model in CLimate Mode)¹⁶ with ERA-Interim-forcing¹⁷ for the period of 1979–2000 were performed for Europe. In all cases, the same model setup, i.e., model version, compiler options, boundary conditions and external parameters was used on all the High Performance Computing (HPC) platforms. The executed codes were identical, and the codes were programmed to produce bitwise-reproducible results for fixed execution environments.

The first ensemble of six simulations with the same initial conditions was run on different platforms, namely, two platforms at Deutsches Klimarechenzentrum (DKRZ) and one platform at Swiss National Supercomputing Centre (CSCS), Leibniz-Rechenzentrum (LRZ), Zentralanstalt für Meteorologie und Geodynamik (ZAMG), and Deutscher Wetterdienst (DWD). The second ensemble of six simulations was run on the same CSCS platform but with different initial states.

To demonstrate the range of variations that emerge in the two ensembles, we show an ensemble of half-year time series of air temperatures (minus ensemble mean) at a height of 2 m at the randomly chosen Oslo-location. Figure 1 shows six different trajectories of the ensembles with different initial value, and the six trajectories of the ensemble results run on different platforms. A visual inspection leads to the conclusion that the two ensembles show the same sequence of divergent and convergent phases, which is confirmed by the time-dependent intra-ensemble standard deviations shown for each day across the trajectories of the two ensembles.

The characteristics of the time series are the same throughout all 20 years. Qualitatively identical results were found for other locations (not shown).

When the standard deviation is large, we can speak of large intra-ensemble variability and an episode of divergence. When the standard deviation is small, we refer to a convergent phase.

A t-test fails to assess the difference of the standard deviations in Fig. 1c as statistically significant. Further, we compared the two ensembles in terms of binned histograms of the standard deviations (Fig. 2). Both histograms of the two ensembles are remarkably similar. This analysis does not explain why the causal effects are identical in both cases. From a computer science perspective, it would be desirable to investigate the effect of changes of infrastructure components on the numerical results. However, such work would require comprehensive statistical analyses, which is beyond the scope of our work here.

At certain times, the unprovoked variability strongly increases, which is the case when the meridional (north-south) gradient of

air pressure (SLP) between 65° N and 45° N averaged from 20° W to 5°W is weak. Such growth is rare when the gradient is strong (not shown). Figure 3 illustrates the difference of the state associated with convergence and with divergence, with values shown as averages across the lowest and highest 5% of standard deviations. How and when the system is disturbed is not important; rather, the presence of the seed or the “numerical noise” is important, and the growth of this disturbance will be initiated if the dynamical state permits.

The emergence of such patterns that favor ensemble divergence has a strong annual cycle. Such divergent phases, such as those measured by local divergence at Oslo, are highest in DJF and smallest in summer JJA (not shown). Further studies of the synoptic states associated with these phases should be performed; however, such work is beyond the scope of this study.

Discussion

First, our experiments confirmed earlier findings, namely, that an ensemble of different trajectories is generated by running the same high-resolution model with slightly different initial conditions. These trajectories are almost identical at certain times but very different at other times⁶. The timing of the episodes of divergence does not depend on the type of disturbance but on the availability of seed noise and on favorable dynamical conditions. Earlier, it was found that if a region is sufficiently small, divergence will rarely occur¹⁸, but with large regions, the frequency and intensity of divergence increases^{16,19}. This finding makes it necessary to frame numerical experiments as stochastic problems.

The additional and new finding here is that the same effect, quantitatively, is obtained by running the same model with identical boundary and initial conditions but on different platforms. The bitwise reproducibility of numerical results can be guaranteed for parallel programs as long as they are executed on the same platform using the same software tools. This reproducibility depends on preserving the order of all mathematical operations and keeping all other components unchanged. In a physical system with limited precision in number representation, the associative property does not hold (for details on number representation, see, e.g.,¹³). For example, $(a + b) + c$ is not necessarily bitwise identical to $a + (b + c)$. The programmer can enforce a reproducible order of execution for all mathematical operations. Even this approach is not trivial^{20,21}. Changing the platform, however, induces hidden changes in the infrastructure that are beyond the control of the programmer. Hence, a different execution order for mathematical operations due to, e.g., different compiler optimizations or library implementations, usually destroys bitwise reproducibility and leads to different simulation results.

Bitwise reproducibility of the results of computer experiments is an important factor for high-quality software development because software errors can be easily detected and the computational validity of scientific results can be checked. When the model is used for numerical experimentation and scenario building, scientists can operate with relaxed requirements. The presence of noise produced by the components of the computer system do not compromise the validity of the scientific results of our experiments. According to our results, this finding holds for changes in the order of execution of mathematical operations. Other approaches suggest to reduce the number of bits for selected variables, i.e., to increase the chance for rounding errors in a controlled way. A benefit of this approach is the lower memory consumption for the code and lower energy consumption during execution because of the higher speed. Also in this case the reproducibility of the simulation results can be enforced.

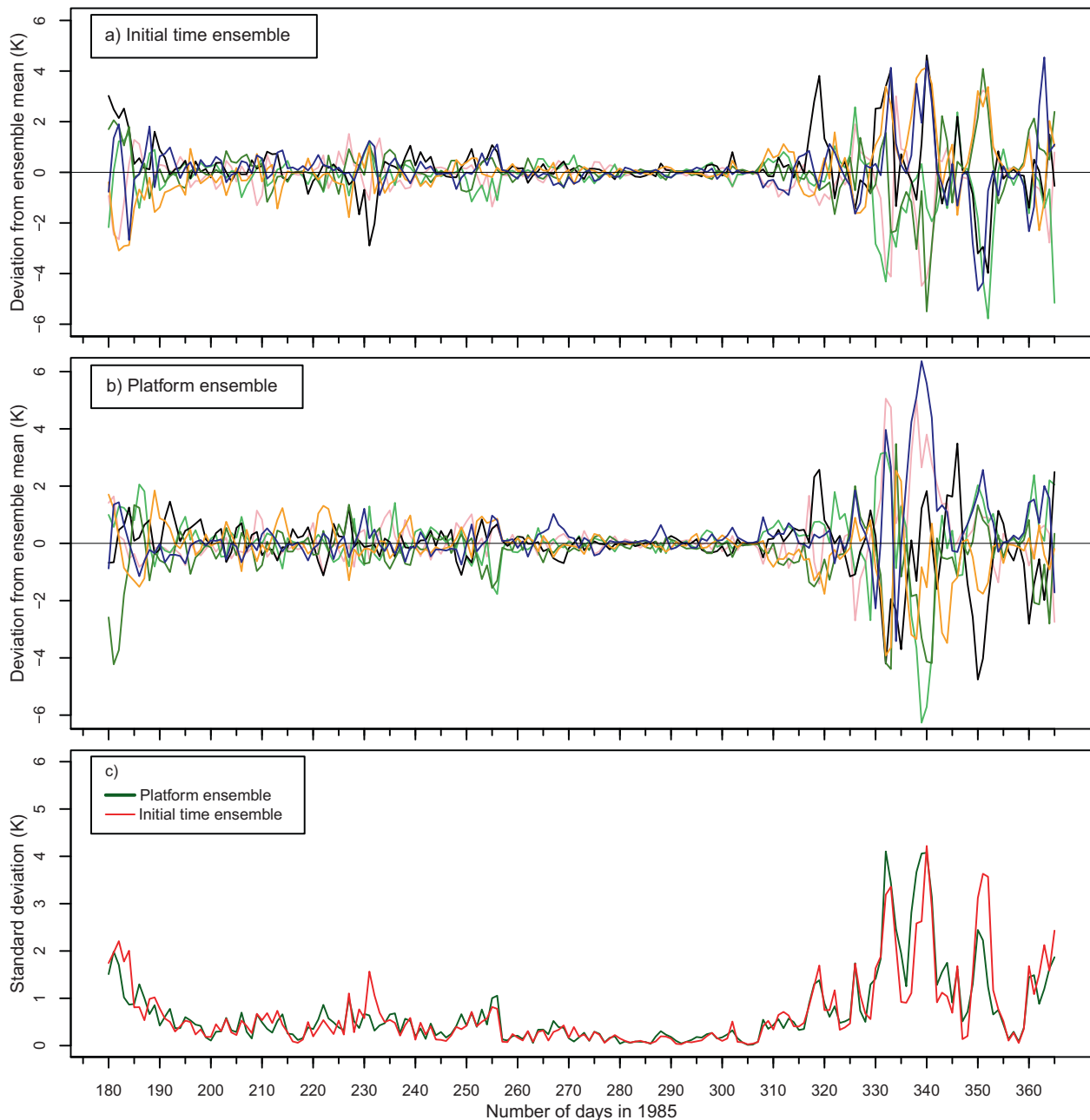


Fig. 1 Air temperature deviations from ensemble mean. **a** Time series of surface air temperatures in the second half of 1985 in the grid cell containing Oslo in six realizations (differently colored) of the initial time ensemble. **b** Time series of surface air temperatures in six realizations (differently colored) of the platform ensemble. **c** Intra-ensemble standard deviations on each day are displayed for both ensembles.

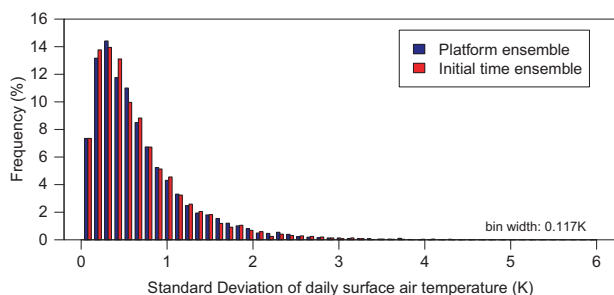


Fig. 2 Binned distributions of 7305 intra-ensemble standard deviations for temperature in Oslo. Histogram of daily intra-ensemble standard deviations in both ensembles (platform ensemble in blue and the initial time ensemble in red) for the daily surface air temperature in Oslo.

However, it requires a thorough analysis of the effect of these changes on the final results. See, e.g., [14,22,23](#).

A possible application involves considering whether the change from a given platform I to another platform II has a significant effect on the numerical results that a given code generates, namely, if platform II generates numbers that are beyond the range of the internal variability of the simulation on platform I. To do so, an ensemble of initial-state-disturbed simulations on platform I is run, and the timing and intensity of the uncertainty are determined. Then, the inter-platform differences can be assessed, and whether these values are within the range of the earlier determined uncertainty can be determined.

We may add some discussion about what “solution” means in this context. We have seen that both types of noise, physical and numerical, lead to different trajectories but to the same

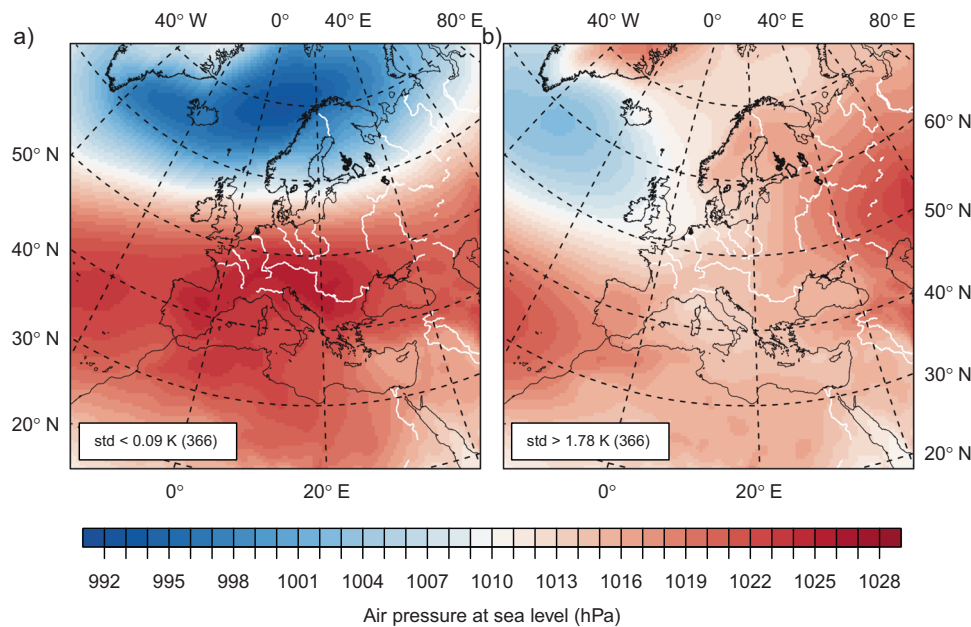


Fig. 3 Surface pressure composites depending on ensemble standard deviation. **a** Composites of mean sea level pressure for states with tendencies toward convergence. The 366 cases with the lowest 5% of standard deviations in the sample are shown. **b** Composites of mean sea level pressure for states with tendencies toward divergence for the initial-state ensemble. The 366 cases with the highest 5% of standard deviations in the sample are shown.

population type of trajectories. Interestingly, it does not matter what type of noise is present; notably, only some minor variations are needed, which under suitable dynamical conditions will lead to significantly large-scale differences in the trajectories.

This is not an error in the code or in the approximation of the dynamics of the system but is a property of the system; i.e., a system can exhibit a number of equally valid but different trajectories. There is not a “true” simulation because at least 5 if not all 6 of the computing platforms considered yield a “false” result. Alternatively, at least 5 or all 6 initial states must be inconsistent with each other to achieve this result, even though they are taken from the observed trajectory and are similar to each other. All six ensemble members are mere random outcomes of a stochastic system. Thus, a “true” solution does not exist.

The concept of a “true” solution is also questionable because it alludes to the existence of “true” differential equations. However, such equations do not exist²⁴. First, the limit $\Delta x \rightarrow 0$ makes little sense. Second, at increasingly small scales, the numbers of processes and state variables increase (e.g., properties of cloud droplets). There is no intrinsic way of modifying the parametrization schemes, which are genuine elements of climate models, when increasing the resolution. The equations encoded in the computer programs operate on grids, and the parametrizations are formulated for these grids. Thus, the equations themselves are mere approximations of an assumed “true” set of equations. All the approximations operate with many degrees of freedom and incorporate various nonlinearities, which lead to what is best conceptualized as stochastic variability, here named “noise”.

Thus, the solution is not to suppress the noise by some numerical measures; instead, this consideration needs to be incorporated into the experimental strategy of the modeler. When asking if a parametrization scheme or different forcing factors induce a change in the probabilistic structure, we need to consider ensembles and compare these ensembles with statistical measures, confidence intervals or hypothesis testing. Alternatively, when the assumptions of stationarity and ergodicity are considered valid, very long simulations may also be considered, with segments of the full record taken as samples. In our case, the latter is not given, as the differences are clearly not stationary.

Methods

The ensembles, initial time and platform were produced using COSMO5.0_clm9. For all ensemble members, the same configuration files had a grid size of 0.44° in rotated coordinates, with 40 vertical levels in terrain-following hybrid height coordinates up to 22.7 km height, 132 pixels in the x direction, 129 pixels in the y direction, and 10 soil levels down to 11.5 m depth. The time step was 300 s with continuous integrations over the entire simulation period; we used a fifth order Runge–Kutta time integration scheme. The COSMO-CLM includes the TERRA-ML scheme²⁵ to parameterize land surface processes, and cumulus convection was parameterized using the Tiedtke scheme²⁶. No spectral nudging was applied.

The external dataset was compiled using GLOBE for orography, GLC2000 for land use, FAO-DSMW for soil parameters and a look-up table²⁵ and MODIS dry and saturated data for the surface albedo with the program EXTPAR v1.6_clm6. The initial and boundary data from ERA-Interim¹⁷ were converted once for all ensemble members to COSMO-CLM input data format with the program int2lm_2.0_clm1. We used temperature, u and v direction of wind velocity, specific cloud liquid water content, specific cloud ice content and specific humidity as 3D fields and surface geopotential, surface pressure, snow surface temperature, surface snow amount, soil temperature, soil moisture and land-sea fraction as single-level fields.

Data availability

During the current study, data from COSMO-CLM simulations were used. The used data are publicly accessible via http://cera-www.dkrz.de/WDCC/ui/Compact.jsp?acronym=DKRZ_LTA_302_ds00003

Code availability

The source code of the used regional climate model COSMO-CLM is available for members of the CLM-Community. The community is open for non-commercial research. It is necessary to register, to sign the terms of use and to contribute to the work of the community.

Received: 20 April 2020; Accepted: 16 December 2020;

Published online: 22 January 2021

References

1. Bocchino, R. L., Adve, V. S., Adve, S. V. & Snir, M. Parallel programming must be deterministic by default. In *Proceedings of the First USENIX Conference on Hot Topics in Parallelism, HotPar09* (USENIX Association, USA, 2009).
2. Chervin, R. M., Gates, W. L. & Schneider, S. H. The effect of time averaging on the noise level of climatological statistics generated by atmospheric general circulation models. *J. Atmos. Sci.* **31**, 2216–2219 (1974).

3. Penduff, T. et al. Chaotic variability of ocean heat content: climate-relevant features and observational implications. *Oceanography* **31**, 63–71 (2018).
4. Ji, Y. & Vernekar, A. D. Simulation of the Asian summer monsoons of 1987 and 1988 with a regional model nested in a global GCM. *J. Climate* **10**, 1965–1979 (1997).
5. Tang, S., von Storch, H. & Chen, X. Atmospherically forced regional ocean simulations of the South China Sea: scale dependency of the signal-to-noise ratio. *J. Phys. Oceanogr.* **50**, 133–144 (2020).
6. Weisse, R., Heyen, H. & von Storch, H. Sensitivity of a regional atmospheric model to a sea statedependent roughness and the need for ensemble calculations. *Monthly Weather Rev.* **128**, 3631–3642 (2000).
7. Gilkeson, C. A. et al. Dealing with numerical noise in CFD-based design optimization. *Comput. Fluids* **94**, 84–97 (2014).
8. Heo, K.-Y. et al. Methods for uncertainty assessment of climate models and model predictions over east asia. *Int. J. Climatol.* **34**, 377–390 (2013).
9. Laurmann, J. A. & Gates, W. L. Statistical considerations in the evaluation of climatic experiments with atmospheric general circulation models. *J. Atmos. Sci.* **34**, 1187–1199 (1977).
10. von Storch, H. & Zwiers, F. *Statistical analysis in climate research* (Cambridge University Press, Cambridge, 1999).
11. Rockel, B., Castro, C., Pielke, R. A., Von Storch, H. & Leoncini, G. Dynamical downscaling: assessment of model system dependent retained and added variability for two different regional climate models. *J. Geophys. Res.* **113**, 1–9 (2008).
12. Hasselmann, K. Stochastic climate models Part I. Theory. *Tellus* **28**: 473–485 (1976).
13. Gustafson, J. L. *The end of error: unum computing* (Chapman and Hall/CRC, 2015).
14. Milroy, D. J. *Climate model quality assurance through consistency testing and error source identification*. Ph.D. thesis (University of Colorado, Boulder, 2018).
15. Milroy, D. J., Baker, A. H., Hammerling, D. M. & 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 (2018).
16. Rockel, B., Will, A. & Hense, A. The regional climate model COSMO-CLM (CCLM). *Meteorologische Zeitschrift* **17**, 347–348 (2008).
17. Dee, D. et al. The era-interim reanalysis: configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.* **137**, 553–597 (2011).
18. Schaaf, B., von Storch, H. & Feser, F. Does spectral nudging have an effect on dynamical downscaling applied in small regional model domains? *Monthly Weather Rev.* **145**, 4303–4311 (2017).
19. Castro, C. L., Pielke, R. A. Sr & Leoncini, G. Dynamical downscaling: assessment of value retained and added using the Regional Atmospheric Modeling System (RAMS). *J. Geophys. Res.* **110**, D05108 (2005).
20. Revol, N. & Théveny, P. Numerical reproducibility and parallel computations: Issues for interval algorithms. *IEEE Trans. Comput.* **63**, 1915–1924 (2014).
21. Mueller, I., Arteaga, A., Hoefler, T. & Alonso, G. Reproducible floating-point aggregation in RDBMSs. *2018 IEEE 34th International Conference on Data Engineering (ICDE)* <https://doi.org/10.1109/icde.2018.00098> (2018).
22. Milroy, D. J., Baker, A. H., Dennis, J. M. & Gettelman, A. Investigating the impact of mixed precision on correctness for a large climate code. In *2019 IEEE/ACM 3rd International Workshop on Software Correctness for HPC Applications (Correctness)* (IEEE, 2019).
23. Tintó Prims, O. et al. 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 (2019).
24. Müller, P. & von Storch, H. *Computer modelling in atmospheric and oceanic sciences - building knowledge* (Springer Verlag Berlin, Heidelberg, New York, 2004).
25. Doms, G. et al. A description of the nonhydrostatic regional COSMO model. Part II: Physical Parameterization. *Consortium for Small-Scale Modelling, Deutscher Wetterdienst* https://doi.org/10.5676/DWD_pub/nwv/cosmo-doc_5.00_II (DWD, 2013).
26. Tiedtke, M. A comprehensive mass flux scheme for cumulus parameterization in large-scale models. *Monthly Weather Rev.* **117**, 1779–1800 (1989).

Acknowledgements

The following computing centers provided the computer hardware for the limited area modelling simulations: German Climate Computing Center (DKRZ), Swiss National Supercomputing Centre (CSCS), Leibniz Supercomputing Centre (LRZ), Zentralanstalt für Meteorologie und Geodynamik (ZAMG), and Deutscher Wetterdienst (DWD). We thank the CLM-Community for assistance, especially H.-J. Panitz (KIT) and B. Rockel (HZG). We thank I. Anders (ZAMG), S. Brien (DWD), K. Keuler (BTU), D. Lüthi (ETHZ), M. Mertens (DLR) for providing their simulation results.

Author contributions

B.G. prepared the experimental set-up, T.L. embedded the issue in the computer science framework, and H.v.S. related the issue to studies on intra-ensemble variability. B.G. and H.v.S. performed the statistical analysis, and all co-prepared the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s43247-020-00085-4>.

Correspondence and requests for materials should be addressed to H.v.S.

Peer review information Primary handling editors: Heike Langenberg.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021