https://doi.org/10.5194/gmd-2021-248
Preprint. Discussion started: 6 September 2021
© Author(s) 2021. CC BY 4.0 License.

Geoscientific Model Development Discussions

An Ensemble-Based Statistical Methodology to Detect Differences in Weather and Climate Model Executables

Christian Zeman[1] and Christoph Schär[1]

[1]Institute for Atmospheric and Climate Science, ETH Zurich, Switzerland

Correspondence: Christian Zeman (christian.zeman@env.ethz.ch)

Abstract. Since their first operational application in the 1950s, atmospheric numerical models have become essential tools in weather and climate prediction. As such, they are subject to continuous changes, thanks to advances in computer systems, numerical methods, more and better observations, and the ever increasing knowledge about the atmosphere of Earth. Many of the changes in today's models relate to seemingly unsuspicious modifications, associated with minor code rearrangements, changes in hardware infrastructure, or software updates. Such changes are meant to preserve the model formulation, yet the verification of such changes is challenged by the chaotic nature of our atmosphere – any small change, even rounding errors, can have a big impact on individual simulations. Overall this represents a serious challenge to a consistent model development and maintenance framework.

Here we propose a new methodology for quantifying and verifying the impacts of minor atmospheric model changes, or its underlying hardware/software system, by using ensemble simulations in combination with a statistical hypothesis test. The methodology can assess effects of model changes on almost any output variable over time, and can also be used with different underlying statistical hypothesis tests.

We present first applications of the methodology with the regional weather and climate model COSMO, including the verification of a major system update of the underlying supercomputer. While providing very robust results, the methodology shows a great sensitivity even to tiny changes. Specific changes considered include the switch from double to single-precision, or the application of a tiny amount of explicit diffusion. Results show that changes are often only detectable during the first hours, which suggests that short-term ensemble simulations (days to months) are best suited for the methodology, even when addressing long-term climate simulations. We also show that the choice of the underlying statistical hypothesis test is not of importance and that the methodology already works well for coarse resolutions, making it computationally inexpensive and therefore an ideal candidate for automated testing.

1 Introduction

Today's weather and climate predictions heavily rely on data produced by atmospheric numerical models. Ever since their first operational application in the 1950s, the models have been improved thanks to advances in computer systems, numerical methods, observational data, and the understanding of the Earth's atmosphere. While such changes often may be only small

https://doi.org/10.5194/gmd-2021-248
Preprint. Discussion started: 6 September 2021
© Author(s) 2021. CC BY 4.0 License.

Geoscientific Model Development Discussions
Open Access
EGU

and incremental, accumulated they have a big effect which manifests itself in a significant increase in skill of weather and climate predictions over the past 40 years (Bauer et al., 2015).

While some changes are there to extend and improve the model, others are not meant to affect the model results but merely its computational performance and versatility. In software engineering, one often makes the distinction between "upgrades" and "updates" in such cases. For weather and climate models, an upgrade would for example be the introduction of a new and improved soil model, whereas a new version of underlying software or a binary that has been built with a newer compiler version would only represent an update. Updates are often employed due to the necessity of keeping the software up-to-date without any perceivable improvements for the user. For a weather and climate model, the model results are not supposed to be significantly affected by such an update. This also applies for other changes, such as the move to a different hardware architecture or changes in the domain decomposition for distributed computing. A robust behavior of the model with regard to such changes is crucial for a consistent interpretation of the results and the credibility of the derived predictions and findings.

Weather and climate model results are generally not bit-identical when they are for example run on different hardware architectures or have been compiled with a different compiler. This is because the associativity property does not hold for floating point operations (i.e. $(x+y)+z = x+(y+z)$ is not given) and the fact that the order of arithmetic operations is dependent on the compiler and the targeted hardware architecture. Schär et al. (2020) have achieved bit-reproducibility for the regional weather and climate model COSMO between a CPU and a GPU version of the model by limiting instruction rearrangements from the compiler and with the use of a preprocessor that automatically adds parentheses to every mathematical expression of the model. However, this also came with a performance penalty where the CPU and GPU bit-reproducible versions were slower by 37% and 13%, respectively, than their non-bit-reproducible counterparts. Due to this performance penalty and the effort involved in making a model bit-reproducible, bit-reproducibility is generally not enforced. It has to be noted that this behavior of not producing bit-identical results across different architectures or with the use of different compilers is common for most computer applications and not a problem per se. But for weather and climate models, it represents a serious challenge due to the chaotic natute of the underlying nonlinear dynamics, where small changes can have a big effect (Lorenz, 1963). For example, a tiny difference in the initial conditions of a weather forecast can potentially lead to a very different prediction. Consequently, also rounding errors can potentially affect the model results in a major way. In order to mitigate this effect and to provide probabilistic predictions, forecasts often use ensemble prediction systems (EPS), where a model is run several times for the same time frame with slightly perturbed initial conditions or stochastic perturbations of the model simulations (see Leutbecher and Palmer, 2008, for an overview). Using an EPS accounts for the uncertainty in initial conditions and the internal variability of the model results.

So in order to verify whether the properties of a weather and climate model executable are not significantly affected after an update or a change to a different platform, we also have to resort to ensemble simulations. Without ensemble simulations, we would only be able to answer something we already know a priori: Any change in the model or its underlying software and hardware will make the model slightly different and therefore most probably also the model output slightly different. With ensemble simulations, we can however answer the much more important question: How do the changes of model results compare to the internal variability of the underlying nonlinear dynamical system? If the effect of the new model is significantly

https://doi.org/10.5194/gmd-2021-248
Preprint. Discussion started: 6 September 2021
© Author(s) 2021. CC BY 4.0 License.

Geoscientific Model Development Discussions

smaller than the one of internal variability, a statistical test will not be able to detect if the results of the new and the old model come from the same distribution or not.

In software engineering terms, verification is a form of system testing. This means that a complete integrated system is tested, in this case a weather and climate model consisting of many different components that interact with each other. Another common form of system testing for weather and climate models is validation. While verification can be defined as a comparison with analytic test cases and computational products, validation is a comparison with observations (Clune and Rood, 2011). In this work, we focus on verification, as we only look at model output. System tests, such as our verification methodology, are an integral part of testing in software engineering. An objective system test, that can be performed automatically, is also a great asset for the practice of continuous integration and continuous deployment (CI/CD). CI/CD enforces automation in building, testing and deployment of applications and should also be considered good practice in the development and operation of weather and climate models.

2 Background

2.1 Current state of the art

Considering its importance for consistency and trustworthiness of model results, verification has received relatively little attention in the weather and climate community. However, the awareness seems to have increased, as especially some recent studies tackle this issue in a more systematic way.

Rosinski and Williamson (1997) were the first to propose a strategy for verifying atmospheric models after they have been ported to a new architecture. They set the conditions that the differences should be within one or two orders of magnitude of machine rounding during the first few timesteps and that the growth of differences should not exceed the growth of initial perturbations at machine precision during the first few days. Thomas et al. (2002) performed 42-hour simulations with the Mesoscale Compressible Community (MC2) model to determine the importance of processor configuration (domain decomposition), floating point precision, and mathematics libraries for the model results. By analyzing the spread of runs with different settings, they concluded that processor configuration is the main contributor among these categories to differences in the results of their dynamical core. Knight et al. (2007) analyzed an ensemble of over 57'000 climate runs from the climateprediction.net project (www.climateprediction.net, last access: 9 June 2021). The climate runs have been performed with varying parameter settings and initial conditions on different hardware and software architectures. With the use of regression tree analysis they demonstrated that the effect of hardware and software is small relative to the effects of parameter variations and, over the wide range of systems tested, may be treated as equivalent to that caused by changes in initial conditions. Hong et al. (2013) performed seasonal simulations with the global model program (GMP) of the Global/Regional Integrated Model system (GRIMs) on 10 different software system platforms with different compilers, parallel libraries, and optimization levels. The results showed that the ensemble spread due to the differences in software system is comparable to the ensemble spread due to the differences in initial conditions.

https://doi.org/10.5194/gmd-2021-248
Preprint. Discussion started: 6 September 2021
© Author(s) 2021. CC BY 4.0 License.

Geoscientific Model Development Discussions

One of the most comprehensive recent studies on the issue of verification is from Baker et al. (2015), where they propose the use of principal component analysis (PCA) for consistency testing of climate models. Instead of testing all model output variables, from which many show high correlations with each other, they determine principal components (PCs) and then compare mean and variances of the first few principal components among ensembles from different configurations (resulting from a code modification, compiler change, or new hardware platform) using z-scores (i.e. if the value from a test configuration is within a certain number of standard deviations from the control ensemble). If the test fails for too many PCs, they reject the new configuration. They confirm their methodology using 1-year long simulations of the Community Earth System Model (CESM) with different parameter settings, hardware architectures, and compiler options. While the methodology shows a high sensitivity and promising results, it has some difficulties in detecting changes at small scales (in their example caused by additional diffusion) due to methodology's focus on annual global mean values. Baker et al. (2016) used the same concept as Baker et al. (2015) for consistency testing of the Parallel Ocean Program (POP), the ocean model component of the Community Earth System Model (CESM). However, instead of evaluating spatial averages, they apply the methodology for each grid point and stipulated that this local test has to pass for at least 90% of the grid points in order to have the global test pass. Milroy et al. (2018) extended the consistency test by Baker et al. (2015) by performing the test on spatial means for the first 9 timesteps of the Community Atmospheric Model (CAM) on a global 1° grid with a timestep of 1800 s. With this method they were able to produce the same results for the same test cases as Baker et al. (2015) and additionally they were also able to detect the diffusion change which was not detected in Baker et al. (2015).

Wan et al. (2017) use timestep convergence as a criterion for model verification, based on the idea that a model executable that is significantly different is no longer expected to converge towards a reference solution that has been produced with the old executable. Their test methodology produces similar results as the one from Baker et al. (2015) and is relatively inexpensive due to the short integration times. However, due to the nature of the test, it is not able to detect issues associated with diagnostic calculations that do not feedback to the model state variables.

Mahajan et al. (2017) used an ensemble-based approach where they applied the Kolmogorov-Smirnov (K-S) test on annual and spatial means of 1-year simulations for testing the equality of distributions of different model simulations. Furthermore, they used generalized extreme value (GEV) theory for representing annual maxima of daily average surface temperature and precipitation rate and then applied a Student's t-test on the estimated GEV parameters at each grid-point in order to test the occurrence of climate extremes. They show that the climate extremes test based on GEV theory is considerably less sensitivity to changes in optimization strategies than the K-S test on mean values. Mahajan et al. (2019) applied two relatively modern multivariate two sample equality of distribution tests, the energy test and the kernel test (see Mahajan et al., 2019, for details on the respective test statistics), on year-long ensemble simulations following Baker et al. (2015) and Mahajan et al. (2017). However, both these tests generally showed a lower power than the K-S test from Mahajan et al. (2017), which means that more ensemble members were needed to confidently reject the null hypothesis.

Massonnet et al. (2020) recently proposed an ensemble-based methodology based on monthly averages (and an average over the whole simulation time), followed by comparison of these averages on a grid-cell level against standard indices used in Reichler and Kim (2008). Finally, spatially averaging results in one scalar number per field, month, and ensemble member. These

https://doi.org/10.5194/gmd-2021-248
Preprint. Discussion started: 6 September 2021
© Author(s) 2021. CC BY 4.0 License.

Geoscientific Model Development Discussions

scalars are then used for the Kolmogorov-Smirnov test (see Sect. 3.3.3) in order to detect statistically significant differences. Performing this test for climate runs with the EC-Earth earth system model version 3.1 on different computing environment revealed significant differences for 4 out of 13 variables. However, the same test for the newer EC-Earth 3.2 version showed no significant differences. Massonnet et al. (2020) suspect the presence of a bug in EC-Earth 3.1 and its subsequent fix for version 3.2 as the reason for this disparity.

2.2 Determining field significance

One statistical approach, that is often used in hypothesis testing for geophysical studies, is the application of local statistical hypothesis tests for each grid cell. Such a statistical hypothesis test is usually performed with a significance level α, which is the probability of rejecting the null hypothesis even though the null hypothesis is true. Let us assume that we have two ensembles, each with n members, and the output on a grid consisting of N grid cells. At each grid point, the hypothesis is tested using an appropriate statistical test and assuming spatial independence. Even if the two ensembles stem from the same model, the test may locally reject the null hypothesis (ensembles stem from same model). When assuming spatial independence, the probability of having x rejected local null hypotheses follows from the binomial distribution:

$$P(x) = \frac{N!}{x!(N-x)!}\alpha^x(1-\alpha)^{N-x} \tag{1}$$

On average, we can expect αN local rejections over the whole grid when two ensembles come from the exact same model. However, the probability of having more than αN rejections is not negligible. For example, for $N = 100$ and $\alpha = 0.05$, the probability of having 9 or more erroneous rejections is still 6.3%, which means that 10 or more local rejections are required (probability 2.8%) in order to reject the global null hypothesis (the model results are indistinguishable) with a 95% confidence interval. This means that for $N = 100$ and $\alpha = 0.05$, 10% of the local hypothesis tests would have to reject the local null hypothesis in order to get a significant global rejection. This percentage is of course much smaller for a larger N. For $N = 10000$, we would require 537 (5.37%) or more local rejections (probability 4.8%) in order to reject the global null hypothesis with a 95% confidence interval (see Fig. 3 in Livezey and Chen, 1983, for a visualization of this function).

However, we have to consider the fact that our local tests cannot be assumed to be statistically independent due to spatial correlation. Therefore, equation 1 is no longer valid. While two identical models will still have αN erroneous rejections on average, having a rather high or low rejection rate becomes more likely due to the spatial correlation. Unfortunately, the exact distribution of rejection rates is unknown in such a case (Storch, 1982). Livezey and Chen (1983) argue that spatial correlation reduces N, the number of independent tests, due to a clustering effect of grid points and therefore also increases the percentage of local rejections needed in order to reject the global null hypothesis. In order to estimate the effective number of independent tests N_{eff}, Livezey and Chen (1983) use Monte Carlo methods by randomly resampling the available data in a way that is consistent with the global null hypothesis. The estimated N_{eff} allows them to again use equation 1 for estimating the number of rejected local tests that are required to reject the global null hypothesis.

https://doi.org/10.5194/gmd-2021-248
Preprint. Discussion started: 6 September 2021
© Author(s) 2021. CC BY 4.0 License.

Geoscientific Model Development Discussions

Wilks (2016) recommends the use of the False Discover Rate (FDR) method by Benjamini and Hochberg (1995). This method defines a threshold level p_{FDR}, based on the sorted p-values. The threshold is defined as

$$p_{\mathrm{FDR}} = \max_{i=1,\ldots,N} \left[p_{(i)} : p_{(i)} \leq (i/N)\alpha_{\mathrm{FDR}} \right], \tag{2}$$

where $p_{(i)}$ are the sorted p-values with $i = 1, \ldots, N$ and α_{FDR} is the chosen control level for the FDR (note that α_{FDR} must not be the same as α for the local test). The FDR method only rejects local null hypotheses, if the respective p-value is no larger than p_{FDR} and thus is supposed to prevent a too high rejection rate due to spatial correlation. The FDR method is computationally much less expensive than Monte Carlo methods and is often used in geophysical studies. However, we found it to be too conservative (i.e. small changes are often not rejected) for our application (not shown here). The rather conservative behavior of the FDR method might be appropriate in climate-change related studies, where the rejection of the global null hypothesis (i.e. "no climate change signal") should be clear and without any doubt. But for our application, the detection of small and often unintentional changes in model behavior, a high sensitivity is desirable. In this application, a high sensitivity could even be interpreted as the conservative approach: After a global rejection (i.e. "the ensemble results produced by the old and new model are not drawn from the same distribution"), one can still investigate and, depending on the magnitude of the effect, accept the change in the model, but be cautious in the future interpretation of the respective fields during the time frame that they have been rejected. Conversely, not rejecting the global null hypothesis despite slight differences could lead to a wrong interpretation and attribution of changes in the model results. We therefore went for a computationally more expensive but also more sensitive approach based on Monte Carlo methods, which is described in section 3.1.

3 Methods and data

3.1 Verification methodology

We consider ensemble simulations of two model versions, which for brevity will be referred to as “old” and “new”, respectively. We start by stating our global null hypothesis:

$H_{0\,(\mathrm{global})}$: The ensemble results from the old and the new model are drawn from the same distribution.

We then consider the changes in the model to be insignificant, if we are not able to reject the global null hypothesis. This global test is based on a statistical hypothesis test applied on a grid-cell level with a local null hypothesis $H_{0\,(i,j)}$. This local null hypothesis is dependent on the used local statistical hypothesis test. We refrain from stating $H_{0\,(i,j)}$ here, as statistical hypothesis tests often have slightly different null hypotheses, as will be shown in Sect. 3.3. It is also important to state that in general, we will not evaluate the whole model output, but compare a limited number of two-dimensional fields, such as the 500 hPa geopotential height or the 850 hPa temperature fields. For each of the fields selected, the two model ensembles will be tested at grid scale against each other, using an appropriate statistical test. The probability of rejecting $H_{0\,(i,j)}$ for two ensembles produced by an identical model is given by the significance level α (here $\alpha = 0.05$). As discussed in Sect. 2.2, the main difficulty of using statistical hypothesis tests on a grid-cell level is the spatial correlation, making the respective tests

https://doi.org/10.5194/gmd-2021-248
Preprint. Discussion started: 6 September 2021
© Author(s) 2021. CC BY 4.0 License.

Geoscientific Model Development Discussions

not statistically independent and thus prohibiting the use of the binomial distribution for calculating the probabilities of grid-averaged rejection rates, when assuming that the two ensembles come from the same model. We chose to deal with this in a conceptually simple but effective way. The methodology follows Livezey (1985): and combines Monte Carlo methods and subsampling to produce a distribution of rejection rates, which can be used to get the probability of having n_{rej} rejections for two ensembles coming from the exact same model. Mahajan et al. (2017) and Mahajan et al. (2019) chose almost the same approach, but they produce the reference distribution by pooling two ensembles (from which they do not know yet whether they come from the same distribution) together and then applying the test to randomly drawn subsets from the pooled ensemble. This approach allows them to bypass the creation of a control ensemble and therefore save compute time. But strictly speaking, the reference value for the number of rejections then comes from a distribution not produced by one but by two models. Depending on the difference between the two models, this might lead to slightly different results compared to a case, where the reference distribution was produced from two identical models. While we assume that the differences between these two approaches will be small in most cases, we still opted for the approach with a control ensemble, since the additionally needed compute time is relatively small for short simulations (see Sect. 3.5).

Figure 1 shows a schematic example of the procedure. The control and reference ensembles come from an identical model (old model), whereas the evaluation ensemble comes from a model where we are not sure whether it produces statistically indistinguishable results (new model). In our case, each ensemble consists of 50 members and we use 100 subsamples consisting of 20 random members drawn from each ensemble (without replacement) in order to calculate the rejection rate distributions. We then test for field significance by comparing the mean rejection rate from the evaluation ensemble to the 0.95 quantile from the control ensemble, rejecting the null hypothesis if the mean rejection rate of the evaluation ensemble is equal or above the 0.95 quantile of the control ensemble rejection rate. It has to be mentioned that the used numbers of ensemble members (50), subsamples (100), and subsample members (20) for this work are a rather arbitrary choice. However, based on some tests with other configurations (not shown in this work), we are quite confident that different choices within a reasonable range will not significantly change the overall behavior of the test.

Next to accounting for spatial correlation, having a rejection rate distribution from a control ensemble also offers more flexibility in evaluating different variables. This will become evident when we look at floored variables, such as precipitation. For precipitation, many grid points of both ensembles will have a zero-value and therefore the test will not be able to reject the null hypothesis even though the two ensembles might come from two very different models. This can lead to a mean rejection rate well below α for two different ensembles and by just looking at α we would conclude that the two ensembles are indistinguishable. However, here we derive the expected rejection rate from the control ensemble and this yields an objective threshold that accounts for such behavior.

It is important to mention that the choice of $\alpha = 0.05$ for the local statistical hypothesis test is arbitrary and does not determine the confidence interval for field significance. Furthermore, the comparison of the mean rejection rate from the evaluation ensemble with the 0.95 quantile from the control might also give a wrong idea of a confidence interval for the field significance. If we assume that the evaluation ensemble comes from an identical model and only take one subsample from the evaluation ensemble, the probability of it having a rejection rate equal or higher than the 0.95 quantile from the

https://doi.org/10.5194/gmd-2021-248
Preprint. Discussion started: 6 September 2021
© Author(s) 2021. CC BY 4.0 License.

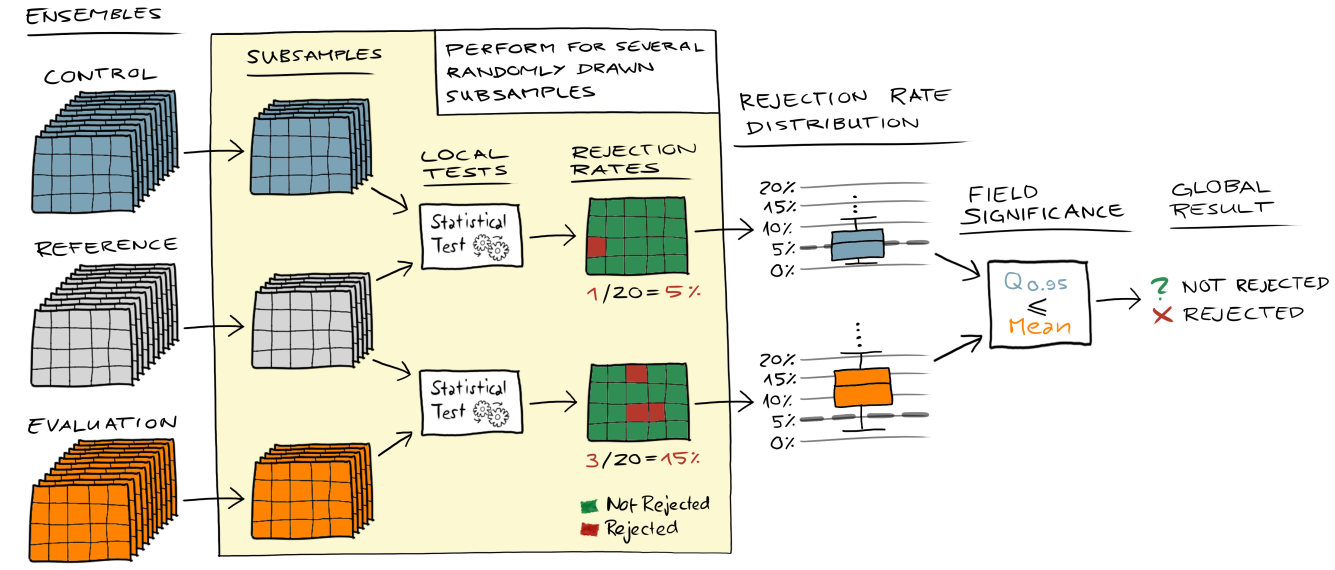

Figure 1. Schematic sketch of the verification methodology. The control and the reference ensemble come from the same "old" model, whereas the evaluation ensemble comes from a "new" model, where we do not know yet whether it is indistinguishable from the model that created the control and reference ensemble. We draw many random subsamples from all three ensembles, perform the local statistical hypothesis tests of the control and evaluation subsamples against the reference subsamples, and then calculate the rejection rate for each subsample. This results in a distribution of the rejection rates for the control and evaluation ensemble, which can then be compared to each other in order to decide whether the evaluation ensemble is different. In this work, we reject the global null hypothesis if the mean of the evaluation ensemble rejection rate distribution is equal or above the 0.95 quantile of the rejection rate distribution from the control ensemble.

control rejection rate distribution is in fact 5%. However, the probability of the mean rejection rate of 100 subsamples from the evaluation ensemble being higher than the 0.95 quantile of the control is significantly lower than 5%, but it is not easy to determine by how much. One could again be tempted to use the binomial distribution in equation 1 for a calculation of the number of necessary rejected subsamples in order to reject the overall null hypothesis. By using the binomial distribution, the probability of rejecting 8 or more subsamples out of 100 would be 6.3% and the probability of rejecting 9 or more subsamples would be 2.8%. So we would need 9 or more rejections to reject the overall null hypothesis with a 95% confidence interval. However, this approach is not fully valid, as the subsamples are not statistically independent from each other. We could again resort to Monte Carlo methods by having a second control and generating a distribution for global rejections of the second control and then use this distribution to define the number of necessary rejection for a given confidence interval. However, this would significantly increase the computational costs. Based on our experience and the results shown in this work, we consider the comparison of the mean to the 0.95 quantile a reasonable choice, even though it is not really based on a confidence interval. But the sensitivity of the methodology could of course be adapted by changing this field significance criterion.

https://doi.org/10.5194/gmd-2021-248
Preprint. Discussion started: 6 September 2021
© Author(s) 2021. CC BY 4.0 License.

Geoscientific Model Development Discussions

The verification methodology in this work shares some similarities with verification methodologies presented in previous studies, most notably Baker et al. (2015, 2016); Milroy et al. (2018); Mahajan et al. (2017, 2019); Massonnet et al. (2020). But most of these studies focus on mean values in space and time. From the previously mentioned studies, only Baker et al. (2016) and Mahajan et al. (2017) have used a similar methodology on a grid cell level, either for monthly averages of variables from an ocean model component (Baker et al., 2016), or for the identification of differences in annual extreme values (Mahajan et al., 2017). And except for Milroy et al. (2018), all other studies focus on longer simulations (one year or more) and average values in time. We will focus on shorter simulations (days to months) with the idea that many small changes are often easier to identify at the beginning of the simulations. We apply the methodology directly to instantaneous or, in the case of precipitation, hourly output variables from an atmospheric model on a 3-hourly or 6-hourly basis. The rejection threshold is computed as a function of time, and may transiently increase or decrease in response to changes in predictability. In essence, the rejection rate distribution from a control ensemble allows us to use an objective criterion for field significance.

3.2 Ensemble generation

The ensemble is created through a perturbation of the initial conditions of the prognostic variables (in our case horizontal and vertical wind components, pressure perturbation, temperature, specific humidity, and cloud water content). The perturbed variable $\hat{\varphi}$ is defined as

$$\hat{\varphi} = (1 + \epsilon R)\varphi\,, \tag{3}$$

where φ is the unperturbed prognostic variable, R a random number with uniform distribution between -1 and 1, and ϵ the specified magnitude of the perturbation. In this study, we have used a magnitude of $\epsilon = 10^{-4}$ for all experiments. This chosen perturbation magnitude proved to be a good compromise between not disturbing the initial conditions too much but still providing a good enough ensemble spread for the statistical verification during the first few hours. Furthermore, choosing such a relatively strong perturbation also allows us to examine the effects of single versus double-precision floating point representation, as the choice minimizes the chance of undesirable rounding artefacts already for the perturbation.

3.3 Statistical hypothesis tests

In this study, we have used three different statistical tests for testing the local null hypothesis $H_{0\,(i,j)}$: the Student's t-test, the Mann-Whitney U test, and the two-sample Kolmogorov-Smirnov test. This allows us to see whether some statistical tests might be better suited for some variables than others and how sensitive the methodology is with regards to the underlying test statistics. If not mentioned otherwise, the Mann-Whitney U test has been used as the default test for the results shown in this study.

3.3.1 Student's t-test

The Student's t-test was introduced by William S. Gosset under the pseudonym "Student" (Student, 1908) and has been originally used to determine the quality of raw material of stout for the Guinness Brewery. The independent two-sample t-test has

https://doi.org/10.5194/gmd-2021-248
Preprint. Discussion started: 6 September 2021
© Author(s) 2021. CC BY 4.0 License.

Geoscientific Model Development Discussions

the null hypothesis that the means of two populations X and Y are equal. As we use it for the local statistical test, we therefore have the following local null hypothesis:

$H_{0\,(i,j)}$: The means $\overline{\varphi_{\text{old}(i,j)}}$ and $\overline{\varphi_{\text{new}(i,j)}}$ are drawn from the same distribution.

Here, $\overline{\varphi_{\text{old}(i,j)}}$ is the sample mean of the variable φ at grid cell (i,j) from the old model, and $\overline{\varphi_{\text{new}(i,j)}}$ is the respective sample mean from the new model. The t statistic is calculated as

$$t = \frac{\overline{X} - \overline{Y}}{s_p \sqrt{\frac{2}{n}}}, \tag{4}$$

with $\overline{X}$ and $\overline{Y}$ being the respective sample means and assuming equal sample size $n = n_X = n_Y$. The pooled standard deviation is given as

$$s_p = \sqrt{\frac{s_X^2 + s_Y^2}{2}}, \tag{5}$$

where s_X^2 and s_Y^2 are the unbiased estimators of the variances of the two samples. The t statistic is then compared against a critical value for a certain significance level α from the Student's t-distribution. For a two-sided test, we reject the local null hypothesis if the t statistic is smaller or greater than this critical value. The Student's t-test requires that the means of the two populations should follow a normal distribution and assumes equal variance. However, the Student's t-test has been shown to be quite robust to violations of both the normality assumption and, provided the sample sizes are equal, the assumption of equal variance (Bartlett, 1935; Posten, 1984). Sullivan and D'Agostino (1992) showed that the Student's t-test even provided meaningful results in the presence of floor effects of the distribution (i.e. where a value can be at minimum zero).

3.3.2 Mann-Whitney U test

The Mann-Whitney U test (also known as Wilcoxon rank-sum test) has been introduced by Mann and Whitney (1947) and is a non-parametric test, in the sense that no assumption is made concerning the distribution of the variables. The null hypothesis is that for randomly selected values X_k and Y_l from two populations, the probability of X_k being greater than Y_l is equal to the probability of Y_l being greater than X_k. It therefore does not test exactly the same property as the Student's t-test (means of two populations are equal), even though it is often compared to it. In our case, the local null hypothesis test for the Mann-Whitney U test is the following:

$H_{0\,(i,j)}$: The probability of $\varphi^k_{\text{old}(i,j)} > \varphi^l_{\text{new}(i,j)}$ is equal to the probability of $\varphi^k_{\text{old}(i,j)} < \varphi^l_{\text{new}(i,j)}$.

Here, $\varphi^k_{\text{old}(i,j)}$ and $\varphi^l_{\text{new}(i,j)}$ are the values of the variable φ at location (i,j) from randomly selected members k and l of the samples from the old and new model respectively. The Mann-Whitney U test ranks all the observations (from both samples combined in one set) and then sums up the ranks of the observations from the respective samples, resulting in R_X and R_Y. U_{min} is calculated as

$$U_{\text{min}} = \min\left(R_X - \frac{n_X(n_X+1)}{2}, R_Y - \frac{n_Y(n_Y+1)}{2}\right), \tag{6}$$

https://doi.org/10.5194/gmd-2021-248
Preprint. Discussion started: 6 September 2021
© Author(s) 2021. CC BY 4.0 License.

Geoscientific Model Development Discussions

where n_X and n_Y are the respective sample sizes. This value is then compared with a critical value U_{crit} from a table for a given significance level α. For larger samples ($n > 20$), U_{crit} is assumed to be normally distributed. If $U_{\text{min}} \leq U_{\text{crit}}$ the null hypothesis is rejected. As a non-parametric test, the Mann-Whitney U test has no strong assumptions and just requires the responses to be ordinal (i.e. <, =, >). Zimmerman (1987) showed that, given equal sample sizes, the Mann-Whitney U test is a bit less powerful than the Student's t-test, even if variances are not equal. This means that the probability of correctly rejecting the null hypothesis, when the alternative hypothesis is true, is assumed to be a bit lower. But when comparing these tests, it is important to keep in mind that they are based on different null hypotheses and thus do not test the same properties.

3.3.3 Two-sample Kolmogorov–Smirnov test

The two-sample Kolmogorov-Smirnov test (hereafter K-S test) is a non-parametric test with the null hypothesis that the samples are drawn from the same distribution. Our local null hypothesis is therefore the following:

$$H_{0\,(i,j)}\colon \quad \varphi_{\text{old}(i,j)} \text{ and } \varphi_{\text{new}(i,j)} \text{ are drawn from the same distribution.}$$

Here, $\varphi_{\text{old}(i,j)}$ and $\varphi_{\text{new}(i,j)}$ are the samples of the variable φ at location (i,j) from the old and new model respectively. The K-S test statistics is given as

$$D_{n_X,n_Y} = \sup_x |F_{X,n_X}(x) - F_{Y,n_Y}(x)|\,, \tag{7}$$

where sup is the supremum function and F_{X,n_X} and F_{Y,n_Y} are the empirical distribution functions of the two samples X and Y, which is defined as

$$F_{X,n_X}(x) = \frac{1}{n_X}\sum_{i=1}^{n_X} I_{[-\infty,x]}(X_i) \tag{8}$$

with the indicator function $I_{[-\infty,x]}(X_i)$, which is equal to one if $X_i \leq x$ and zero otherwise. The null hypothesis is rejected if

$$D_{n_X,n_Y} > c(\alpha)\sqrt{\frac{n_X + n_Y}{n_X \cdot n_Y}}\,, \tag{9}$$

where $c(\alpha) = \sqrt{-\ln(\frac{\alpha}{2}) \cdot \frac{1}{2}}$ for a given significance level α. The K-S test is often perceived to be not as powerful as for example the Student's t-test for comparing means and measures of location in general (Wilcox, 1997). However, due to its different null hypothesis, it might be a more suitable test for testing the shape or the spread of a distribution.

3.4 Model description and hardware

The Consortium for Small-scale Modelling (COSMO) model (Baldauf et al., 2011) is a regional model which operates on a grid with rotated latutide-longitude coordinates. It has been originally developed for numerical weather prediction, but has been extended to also run in climate mode (Rockel et al., 2008). COSMO uses a split explicit third-order Runge-Kutta discretization (Wicker and Skamarock, 2002) in combination with a fifth-order upwind scheme for horizontal advection, and an implicit

https://doi.org/10.5194/gmd-2021-248
Preprint. Discussion started: 6 September 2021
© Author(s) 2021. CC BY 4.0 License.

Geoscientific Model Development Discussions

Crank-Nicholson scheme for vertical advection. Paremeterizations include a radiation scheme based on the δ-two-stream approach (Ritter and Geleyn, 1992), a single-moment cloud microphysics scheme (Reinhardt and Seifert, 2006), a turbulent kinetic energy based parameterization for the planetary boundary layer (Raschendorfer, 2001), an adapted version of the convection scheme by Tiedtke (1989), a subgrid-scale orgogaphy (SSO) scheme by Lott and Miller (1997), and a multi-layer soil model with a representation of groundwater (Schlemmer et al., 2018). Explicit horizontal diffusion is applied by using a monotonic 4th-order linear scheme acting on model levels for wind, temperature, pressure, specific humidity, and cloud water content (Doms and Baldauf, 2018) with an orographic limiter which helps avoiding excessive vertical mixing around mountains. For the standard experiments in this paper, the explicit diffusion from the monotonic 4th-order linear scheme is set to zero.

Most experiments in this work have been carried out with version 5.09. While COSMO has been originally designed to run on CPU architectures, this version is also able to run on hybrid GPU-CPU architectures thanks to an implementation described in Fuhrer et al. (2014), which was a joint effort from MeteoSwiss, the ETH-based Center for Climate Systems Modeling (C2SM), and the Swiss National Supercomputing Center (CSCS). The implementation makes use of the domain-specific language GridTools for the dynamical core, and OpenACC compiler directives for the parameterization package. The simulations have been carried out on the Piz Daint supercomputer at CSCS, using Cray XC50 compute nodes consisting of a Intel Xeon E5-2690 v3 CPU and a NVIDIA Tesla P100 GPU. Except for one ensemble that has been created with a COSMO binary that exclusively uses CPUs, all simulations in this paper have been run in hybrid GPU-CPU mode where the main load of the work is done by the GPUs.

3.5 Domain and Setup

The domains that have been used for the simulation and verification includes most of Europe and some part of Northern Africa (see Fig. 2). The simulated periods all start on 28 May 2018 at 00:00 UTC and range from several days to 3 months in length. The initial and the 6-hourly boundary conditions come from from the European Centre for Medium-Range Weather Forecasts (ECMWF) ERA-Interim reanalysis (Dee et al., 2011). For this work, we have chosen a $132 \times 129 \times 40$ grid with 50 km horizontal grid spacing and the 40 non-equidistant vertical levels reaching up to a height of 22.7 km. In order to reduce the effect of the lateral boundary conditions, we excluded 15 grid points at each of the lateral boundaries from the verification, resulting in 102×99 grid points for one vertical layer. As the verification methodology is supposed to be used as a part of an automated testing environment, we have chosen this relatively coarse resolution in order to keep the computational and storage costs low. Running such a simulation for 10 days requires about 4 minutes on one Cray XC50 compute node when using the GPU-accelerated version of COSMO in double-precision. This means that an ensemble of 50 members requires 3 to 4 node hours and, as the runs can be executed in parallel, the generation of the ensemble is only a matter of minutes.

https://doi.org/10.5194/gmd-2021-248
Preprint. Discussion started: 6 September 2021
© Author(s) 2021. CC BY 4.0 License.

Geoscientific Model Development Discussions

3.6 Experiments

In order to test and demonstrate the methodology, we have performed a series of experiments. Many of these experiments are for cases where we deliberately changed something on the model. However, we also have one real-world case where we verified the effect of a major update of the supercomputer Piz Daint, on which we have been running our model.

3.6.1 Diffusion experiment

COSMO offers the possibility of applying explicit diffusion with a monotonic 4th-order linear scheme with orographic limiter acting on model levels for wind, temperature, pressure, specific humidity, and cloud water content. Diffusion is applied by introducing an additional operator at the right hand side of the prognostic equation, similar to

$$\frac{\partial \psi}{\partial t} = S(\psi) + D \cdot c_d \cdot \nabla^4 \psi \,, \tag{10}$$

where ψ is the prognostic variable, S represents all physical and dynamical source terms for ψ, c_d is the default diffusion coefficient in the model, and D is the factor that can be set in order to change the strength of the computational mixing (please refer to Seciton 5.2 in Doms and Baldauf, 2018, for the exact equations including the limiter). By default, we have set $D = 0$, which means that no explicit 4th-order linear diffusion is applied. However, for one experiment we have set $D = 0.01$. Such a small value of 0.01 should not affect the model results in a way that is visible by eye or easily quantifiable without statistical testing. For a comparison, Zeman et al. (2021) have used values as high as $D = 4.0$ for a model intercomparison between COSMO and the Integrated Forecast System (IFS) from ECMWF. For such high values, the model results visibly change.

3.6.2 Architecture: CPU vs GPU

Per default, the simulations shown in this work have been performed with a binary which makes use of the NVIDIA Tesla P100 GPU on the Cray XC50 nodes (see Section 3.4 for details). For this experiment we have produced an ensemble from the identical source and with identical settings, but compiled to run exclusively on the Intel Xeon E5-2690 v3 CPUs in order to see whether there is a noticeable difference between the CPU version and the GPU version of COSMO.

3.6.3 Floating point precision

In this work, COSMO has been using the double-precision (DP) floating point format by default, where the representation of a floating point number requires 64 bits. However, COSMO can also be run in single-precision (SP) floating point representation. The single-precision version has been developed by MeteoSwiss and is currently used by them for their operational forecasts. They have decided to use the single-precision version after having carefully evaluated its performance compared to the double-precisison version, which suggests that there are only very small differences. However, a reduction of precision leads to greater round-off errors and thus could lead to a noticeable change in model behavior. In order to see whether our methodology would be able to detect differences, we have applied it for a case where the evaluation ensemble has been produced by the single-precision version of COSMO and the control and reference ensembles by the double-precision version of COSMO. It has to

https://doi.org/10.5194/gmd-2021-248
Preprint. Discussion started: 6 September 2021
© Author(s) 2021. CC BY 4.0 License.

Geoscientific Model Development Discussions

be mentioned that for the single-precision version of COSMO, the soil model and parts of the radiation model are still using double-precision, as some discrepancies were detected during the development of the single-precision version.

Running COSMO on one node in single-precision, where a floating point number only requires 32 bits, gives a speedup of around 1.1 for our simulations, which is most likely due to the increased operational intensity (number of floating point operations per number of bytes transferred between cache and memory). When running on more than one node, it is often possible to reduce the total number of nodes for the same setup when switching to single-precision, thanks to a drastic reduction of required memory. For example, a model domain and resolution that usually requires four nodes in double-precision (e.g. the same domain as in this paper, but with 12 km grid spacing instead of 50 km grid spacing), often only requires two nodes in single-precision. This results in a coarser domain decomposition and thus less overlapping grid cells whose values have to be exchanged between the nodes. Combined with the reduced number of bytes of the floating point values that have to be exchanged, a significant reduction of data transfer via the interconnect can be achieved, which increases the efficiency of the system. While running in single-precision on only two nodes might be slower than running the same simulation in double-precision on four nodes, it requires much less node hours. In this particular case (4 nodes for DP vs. 2 nodes for SP), the speedup in node hours was around 1.4, which makes the use of single-precision an attractive option.

3.6.4 Vertical heat diffusion coefficient and soil effects

In order to test the methodology for slow processes related to the hydrological cycle, we have set up an experiment where we induce a relatively small but still notable change. One parameter that has been deemed important to the COSMO model calibration by Bellprat et al. (2016) is the minimal diffusion coefficient for vertical scalar heat transport *tkhmin*. It basically sets a lower bound for the respective coefficient used in the 1D turbulent kinetic energy (TKE) based subgrid-scale turbulence scheme (Doms et al., 2018). By default, we have used a value of $tkhmin = 0.35$ for our simulations, but for this evaluation ensemble we have changed it to $tkhmin = 0.3$. This is not a huge change, as for example the default value in COSMO is set to $tkhmin = 1.0$, whereas the German Weather Service (DWD, Deutscher Wetter Dienst) uses $tkhmin = 0.4$ for their operational model with 2.8 km grid spacing (Schättler et al., 2018). The goal of this experiment is to see whether such a change becomes detectable in the slowly changing soil moisture variable, and if yes, how long it takes to propagate the signal through the different soil layers.

3.6.5 No subgrid-scale orography parameterization

So far, the experiments have been set up for cases where there are only slight model changes. In order to see whether the methodology is able to confidently reject results from significantly different models, we have applied it on an evaluation ensemble where the model had the subgrid-scale orography (SSO) paramterization by Lott and Miller (1997) switched off. At a grid spacing of 50 km, orography cannot be realistically represented in a model, which is why the parameterization should be switched on in order to account for orographic form drag and gravity wave drag effects. Zadra et al. (2003) and Sandu et al. (2013) both showed improvements in both short- and medium-range forecasts with a SSO parameterization based on the formulation by Lott and Miller (1997) for the Canadian Global Environmental Mutiscale (GEM) model and the ECMWF

https://doi.org/10.5194/gmd-2021-248
Preprint. Discussion started: 6 September 2021
© Author(s) 2021. CC BY 4.0 License.

Geoscientific Model Development Discussions

Integrated Forecast System (IFS). Pithan et al. (2015) showed that the parameterization was able to significantly reduce biases in large-scale pressure gradients and zonal wind speeds in climate runs with the general circulation model ECHAM6. So we expect the test to clearly reject the global null hypothesis within the first few days, but also for a longer period of time, which is why we use model runs of 90 days for this experiment.

3.6.6 Piz Daint Update

The supercomputer Piz Daint at the Swiss National Supercomputing Center (CSCS), has recently received two major updates on 9 September 2020 and 16 March 2021. The major changes that affected COSMO were new versions of the Cray Programming Toolkit (CDT) which changed the compilation environment for COSMO, with the new version being CDT 20.08 compared to the old version CDT 19.08 before the first update in September 2020. Both changes were associated with the loss of bit-identical execution. Using containers, CSCS was able to create a testing environment for us that replicated the environment before the first update on 9 September 2020 with CDT 19.08. With this environment we were able to reproduce the results from runs before the update in a bit-identical way. So by using this containerized version and comparing its output to the output from the executable that has been compiled in the updated environment with CDT 20.08, we were able to apply our methodology for a realistic scenario. With the term "realistic" we mean that the system update addressed changes that are typical in a model development context. Indeed, the system upgrade of the Piz Daint software environment was the motivation for the current study.

4 Results

4.1 Diffusion experiment

Here, we discuss the results from the diffusion experiment described in Sect. 3.6.1. Figure 2 shows, why it is important to have such a statistical approach for verification. By just looking at the mean values of the ensembles and their differences (in this case 850 hPa temperature), it is impossible to say whether the two ensembles come from the same distribution. There are some small differences, but these could also just be a product of internal variability and the tiny amount of additional explicit diffusion in the diffusion ensemble is not visible by eye. However, the mean rejection rates calculated with the methodology are clearly higher for the diffusion ensemble in some places in comparison to the control, indicating that the ensembles do not come from the same model. This becomes clear when we compare the mean rejection rate for 500 hPa geopotential of the diffusion ensemble to the 0.95 quantile of the control at the bottom of Fig. 3. The methodology is able to reject the global null hypothesis for the first 60 hours. Afterwards, it is no longer able to reject it, which indicates that from this point on the effect of internal variability is greater than the one from the additional explicit diffusion.

In Fig. 3, we can also see that the mean rejection rate of the control is very close the the expected 5%, which is the significance level α of the underlying Mann-Whitney U test. However, the rejection rate of some samples in the control deviate by quite much from 5% even though the results come from an identical model. Generally, the spread or rejection rates also becomes

https://doi.org/10.5194/gmd-2021-248
Preprint. Discussion started: 6 September 2021
© Author(s) 2021. CC BY 4.0 License.

Geoscientific Model Development Discussions

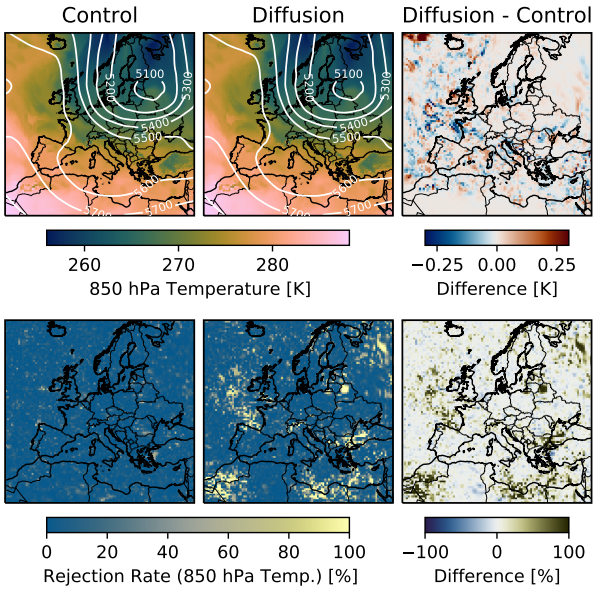

Figure 2. The top row shows the ensemble-mean 850 hPa temperature (color shading) and 500 hPa geopotential height (white contours) for the control (left) and diffusion ensemble (middle) after 24 hours, using all 50 members per ensemble. The difference in mean temperature is shown in the top right panel. The bottom row shows the mean rejection rate for 850 hPa temperature (calculated with the Mann-Whitney U test for 100 subsamples with 20 members per subsample) for each grid cell for these two ensembles, as well as their difference. The substantial differences in the mean rejection rates indicate clearly that the two ensembles come from different models.

bigger with time, which likely is related to changes in spatial correlation and/or decreasing predictability. While the initial perturbations are random and therefore not spatially correlated, the statistical independence becomes already invalid after the first timestep, as a perturbation of a value in a grid cell will naturally affect the corresponding values in the neighboring grid cells. This increasing spread emphasizes the importance of having such a control rejection rate for the decision on the evaluation ensemble.

4.2 Architecture: CPU vs GPU

The COSMO executable running on CPUs does not lead to any global rejections compared to the executable running mainly on GPUs, which is exemplified in Fig. 3 for 500 hPa geopotential. So while the results are not be bit-identical, we consider the difference between these two executables negligible. This confirms that the GPU implementation of the COSMO model is of very high quality, as in terms of execution it cannot be distinguished from the original CPU implementation. This bespeaks an impressive achievement given that the whole code (dynamical core and parameterization package) had to be refactored.

https://doi.org/10.5194/gmd-2021-248
Preprint. Discussion started: 6 September 2021
© Author(s) 2021. CC BY 4.0 License.

Geoscientific Model Development Discussions
Open Access
EGU

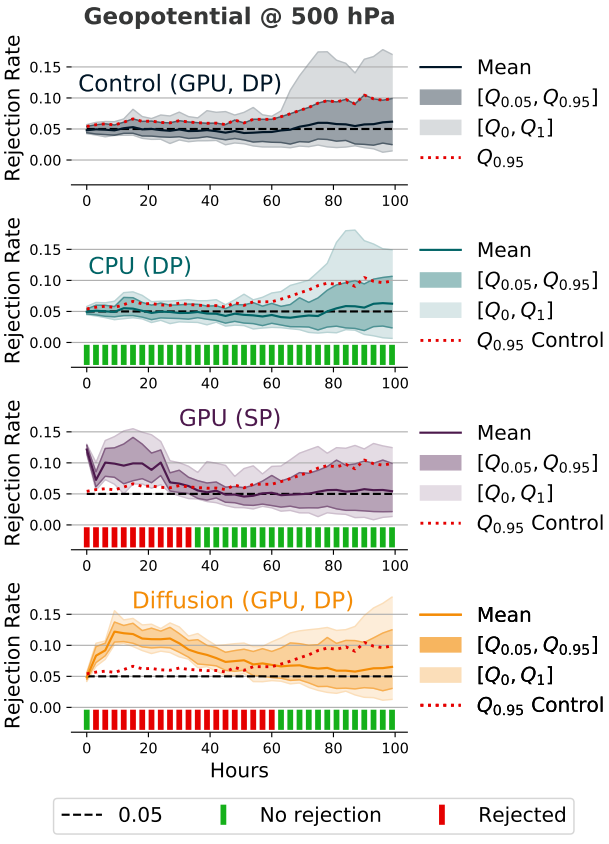

Figure 3. Rejection rates and decisions for 500 hPa geopotential using the Mann-Whitney U test as a underlying statistical hypothesis test. The reference and control ensemble were produced by COSMO running on GPUs in double-precision (top) and it was compared against (from top to bottom) COSMO running on CPUs in double-precision, on GPUs in single-precision, and on GPUs in double-precision with additional explicit diffusion ($D = 0.01$). We reject the null hypothesis if the mean rejection rate is above the 95th percentile of the rejection rate distribution from the control ensemble (red dotted line). The test detects no differences for the CPU version in double-precision, but it detects differences for the other two ensembles during the first few hours/days. The rejections for the initial conditions of the single-precision is most likely associated with differences in the diagnostic calculation of the geopotential due to the reduced precisison.

4.3 Floating point precision

The results of the verification of the single-precision version of COSMO against the corresponding double-precision can be seen in Fig. 3 for the 500 hPa geopotential. Before discussing the results, we remind the reader that some of the variables, notably in the soil model and the radiation codes, are retained in the double-precision version, as some discrepancies were detected during the development of the single-precision version. It should be noted that the geopotential is a plain diagnostic field in the COSMO model, so it is not perturbed initially, but diagnosed at output time from the prognostic variables. However, as the geopotential is vertically integrated, it encompasses information from many levels and variables, and can thus be considered

https://doi.org/10.5194/gmd-2021-248
Preprint. Discussion started: 6 September 2021
© Author(s) 2021. CC BY 4.0 License.

Geoscientific Model Development Discussions
Open Access
EGU

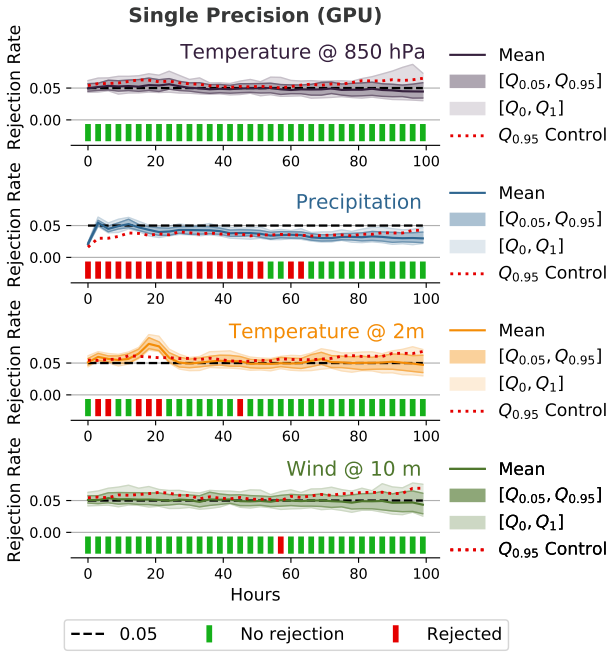

Figure 4. Rejection rates and decisions as in Fig. 3, but for additional variables of the single-precision ensemble. The rejection of precipitation during the first 60 hours suggests that there are also differences between SP and DP associated with the microphysics, whereas the less clear rejection of 2 m temperature might be an effect of the difference in precipitation. Other variables, such as 850 hPa temperature or 10 m wind, show no significant differences.

a well-suited field for testing. One of the most striking features in Fig. 3 is that the methodology rejects the single-precision version already at the initial state of the models. At this state, the perturbation has already been applied according to equation 3, but the model has not yet started the integration in time, which is why we for example see no rejection for the ensemble with additional explicit diffusion in the same plot. It is not entirely clear why the rejection rate is that high for the initial conditions, but we assume that there is a small difference in the calculation of the 500 hPa geopotential due to increased roundoff errors for the vertical integration. Considering that the small perturbations did not yet have time to grow, there is no real internal variability which could "hide" that difference. After 3 hours, the mean rejection rate of the single-precision ensemble is substantially lower, but still higher than the 0.95 quantile from control. At this point in time, the rejection rate is similar to the one from the diffusion example and follows a similar trajectory. In order to rule out differences in perturbation strength due to rounding errors (see also Sect. 3.2), we have performed the same experiment for a modified double-precision version of COSMO, where the to-be perturbed fields are casted to single-precision, the perturbation is applied in single-precision, and the fields are then casted back to double-precision. However, this had no effect on the results and the single-precision ensemble was still rejected with the same magnitude for the initial conditions. We do not see such a clear initial rejection for other variables, as can be seen in Fig. 4. This strengthens our assumption that this initial difference comes from the increased roundoff in the diagnostic calculation

https://doi.org/10.5194/gmd-2021-248
Preprint. Discussion started: 6 September 2021
© Author(s) 2021. CC BY 4.0 License.

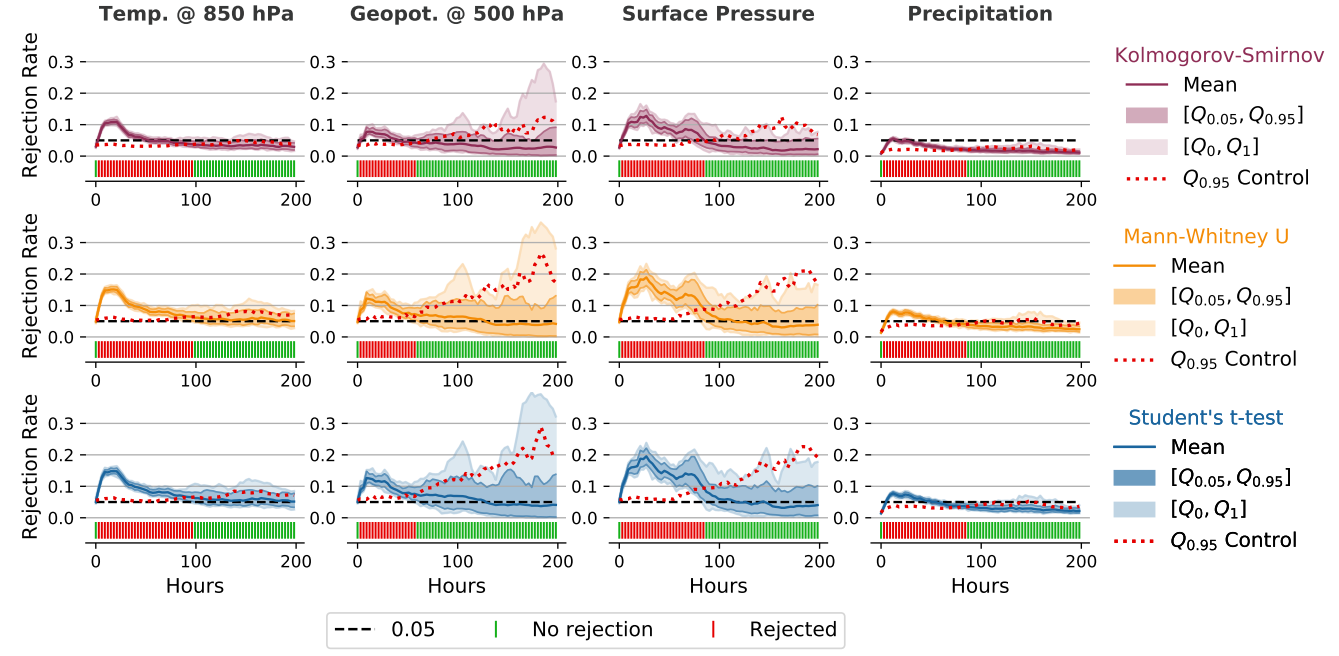

Figure 5. Rejection rates and decisions similar to Fig. 3 for different variables and with the use of different underlying statistical hypothesis tests for a model configuration with additional explicit diffusion. While the rejection rates show some differences amongst the used variables, the decisions are very similar throughout all variables and tests. For most combinations of tests and variables, the methodology is no longer able to detect a difference after around 100 hours. For the 500 hPa geopotential, the methodology is not able to detect differences already after 60 hours due to the high variability of individual rejection rates for this variable. The rejection rates with the Kolmogorov-Smirnov test are usually a bit lower than for the other two tests, but this has no effect for the global decisions, as the respective 0.95 quantiles from the control ensemble are also lower. The Student's t-test shows very similar rejection rates as the non-parametric Mann-Whitney U test, even for variables such as precipitation, which is clearly not normally-distributed.

of the 500 hPa geopotential. The rejection of precipitation during the first 50 to 65 hours in Fig. 4 indicates that there are also some differences associated with the microphysics between single-precision and double-precision. The less clear rejection of 2 m temperature might be a consequence of differences in precipitation. For other variables, such as 850 Pa temperature or 10 m wind, the methodology is not really able to detect differences, which shows that the reduced precision does not affect all variables to the same extent.

4.4 Statistical hypothesis tests

We have tested our methodology with the different statistical hypothesis tests described in Sect. 3.3 for the test case with additional explicit diffusion (see above). Figure 5 shows the respective rejection rates and decisions for several variables. For all variables shown here, the rejection rates from the Student's t-test and the Mann-Whitney U test are almost identical.

https://doi.org/10.5194/gmd-2021-248
Preprint. Discussion started: 6 September 2021
© Author(s) 2021. CC BY 4.0 License.

Geoscientific Model Development Discussions

This confirms the robust behavior of the Student's t-test, despite violations of the normality assumptions and floor effects. This is especially exemplified by the results for precipitation, where the means of the distribution do not follow a normal distribution and also are floored (no negative precipitation). Like the Mann-Whitney U test, the Kolmogorov-Smirnov test is a non-parametric test and therefore does not rely on assumptions about the distribution of the variables. However, its rejection rate is generally a bit lower than that of the Mann-Whitney U test and the Student's t-test. This effect can also be seen in the 0.95 quantile of the control rejection rate, which is generally a bit lower than for the other two hypothesis tests. This is most likely associated with the lower power of the Kolmogorov-Smirnov test (see Sect. 3.3.3). However, the decision (reject or not reject) is always the same in this case for all tests. This indicates that any of these tests is suitable as an underlying statistical hypothesis test and that the choice of the statistical test is not very critical for our methodology. Nevertheless, we have decided to use the Mann-Whitney U test for the subsequent experiments shown in this work, as it offers a slightly higher rejection rate than the Kolmogorov-Smirnov test and, as a non-parametric test, its use is easier to justify than the use of the Student's t-test, even though these two produce almost identical results.

Another interesting aspect is that for three out of four variables in Fig. 5, the methodology rejects the global null hypothesis for close to the first 100 hours. However, as already shown in Sect. 4.1, the methodology is no longer able to reject the global null hypothesis for the 500 hPa geopotential after 60 hours. Why exactly this is the case is not clear and would require further studies, but the result shows that a model change will not affect all variables the same way. Consequently, it is advisable to perform the test for a set of variables instead of only one variable. Baker et al. (2015) first perform principal component analysis (PCA) on the model output variables and then apply their consistency test only on the first few principal components (PCs) instead of 120 variables. As many of the variables are highly correlated, this approach lets them represent most of the variance in the data in only a few PCs. This might also be an option for the verification methodology shown in this work. While the rather abstract nature of a PC makes it more difficult to attribute a rejection to a certain process, the verification could still be performed directly on specific output variables after a rejection in order to investigate the difference in more detail.

4.5 Vertical heat diffusion and soil effects

Figure 6 shows the rejections rates and global decisions for the 2 m temperature and soil moisture at different depths for the model setting with a modified minimal diffusion coefficient for vertical scalar heat transport (*tkhmin* = 0.3 instead of 0.35). Note that this change will only be effective at a subset of the gridpoints (as *tkhmin* represents a limiter). The rejection rate is quite high for the 2 m temperature during the first few days. For the soil moisture at different depths, we can see that the magnitude of the rejection rate decreases the deeper we go. Furthermore, the initial perturbation and the subsequent internal variability of the atmosphere clearly need some time to travel to the lower layers which is most obvious in the layer at 2.86 m depth. In this layer, the rejection rate remains close to zero for the first few days because there is almost no difference visible between the different ensemble members. As a consequence of the not yet "arrived" perturbation, the global decision for this layer should be interpreted with caution during these first few days. But while the magnitude and the variability of the rejection rate is decreasing for the lower soil layers, the effect is visible for longer, which is clearly related to the slower processes in the soil. For 2 m temperature, there are still some rejections after 50-60 days. However, the test is usually not able to reject the

https://doi.org/10.5194/gmd-2021-248
Preprint. Discussion started: 6 September 2021
© Author(s) 2021. CC BY 4.0 License.

Geoscientific Model Development Discussions

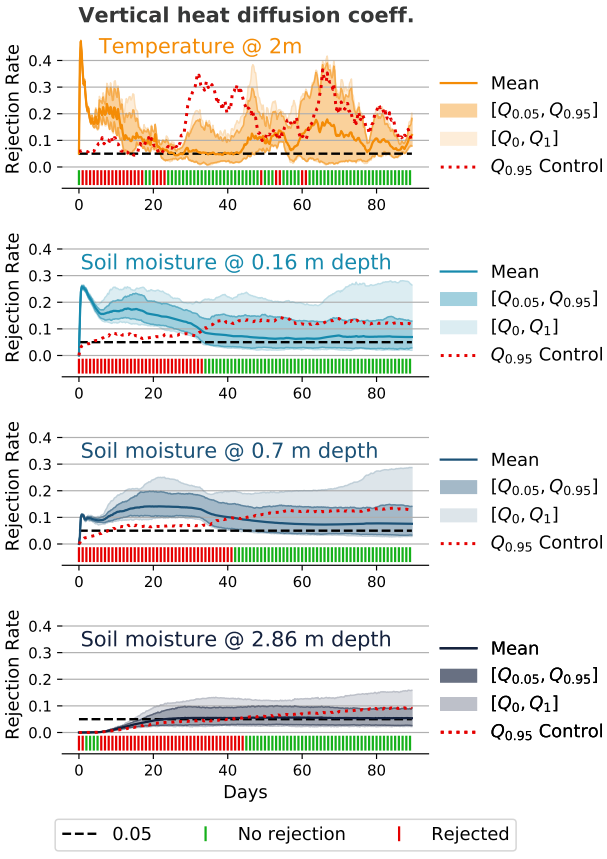

Figure 6. Rejection rates and decisions similar to Fig. 3 for 2 m temperature and soil moisture at different depths for an ensemble where the the minimal diffusion coefficient for vertical scalar heat transport has been slightly changed (*tkhmin* = 0.3 instead of 0.35). The initial random perturbation of the atmosphere clearly needs some time to travel to the deeper soil layers. Furthermore, the magnitude of the rejection rate clearly is lower for the deeper soil layers, but the difference is noticeable for a longer time period.

global null hypothesis for 2 m temperature after 25 days, which indicates that from this point on the effect from the change of *tkhmin* has been overshadowed by internal variability.

4.6 No subgrid-scale orography parameterization

Disabling the SSO parameterization is an important change and our methodology is able to clearly detect this for the whole 3 months simulation time. The mean rejection rate for the three variables shown in Fig. 7 is very high and seems to remain at a relatively constant level after the first month. This indicates that the difference would also be detectable after a longer simulation time, even though the variability on a grid cell level must be very high.

https://doi.org/10.5194/gmd-2021-248
Preprint. Discussion started: 6 September 2021
© Author(s) 2021. CC BY 4.0 License.

Geoscientific Model Development Discussions

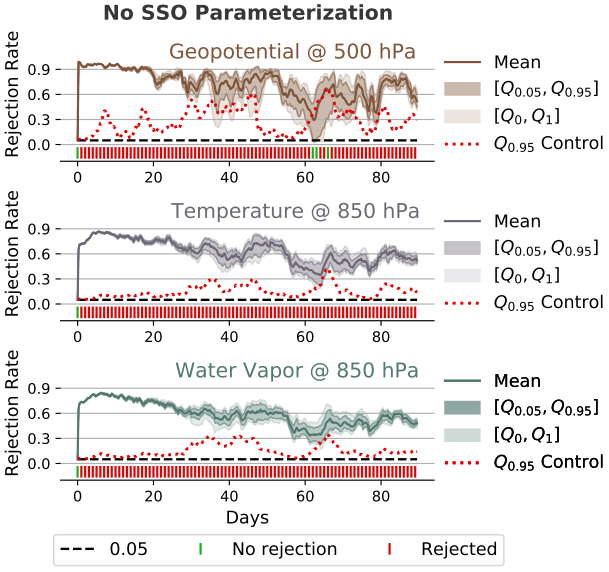

Figure 7. Rejection rates and decisions similar to Fig. 3 but for the 500 hPa geopotential, 850 hPa temperature, and 850 hPa water vapor amount and for an evaluation ensemble where the subgrid-scale orography (SSO) parameterization has been switched off. The methodology clearly rejects the null hypothesis throughout all 90 days, except in three instances for the 500 hPa geopotential. The difference between the mean rejection rate of the evaluation ensemble and the 0.95 quantile of the control is quite large and persistent, which indicates that such a big change in the model is detectable for an even longer time.

4.7 Piz Daint Update

Figure 8 shows that we do not detect any differences after the update of the supercomputer Piz Daint. Considering how closely the 0.95 quantile from the control ensemble follows the 0.95 quantile from the evaluation ensemble and how close the mean rejection rate from the evaluation ensemble is to 0.05, even increasing the sensitivity of the test by using a lower quantile for the determination of field significance would most probably not change the result. Therefore, we are confident that the system update did not affect the model in any significant way.

5 Discussion about applicability

As opposed to the existing verification methodologies described in Sect. 2, our methodology does not rely on any averaging in space and/or time. This approach offers several advantages. The verification on a grid-cell level allows us to identify differences in small-scale and short-lived features that maybe do not affect spatial or temporal averages. Furthermore, it provides fine-grained information in space and time and therefore represents already useful information for the debugging process. A good example for this is the initial rejection of the 500 hPa geopotential field for the single-precision experiment. The test rejects the null hypothesis already before the integration in time has started, which indicates that there are already detectable differences

https://doi.org/10.5194/gmd-2021-248
Preprint. Discussion started: 6 September 2021
© Author(s) 2021. CC BY 4.0 License.

Geoscientific Model Development Discussions

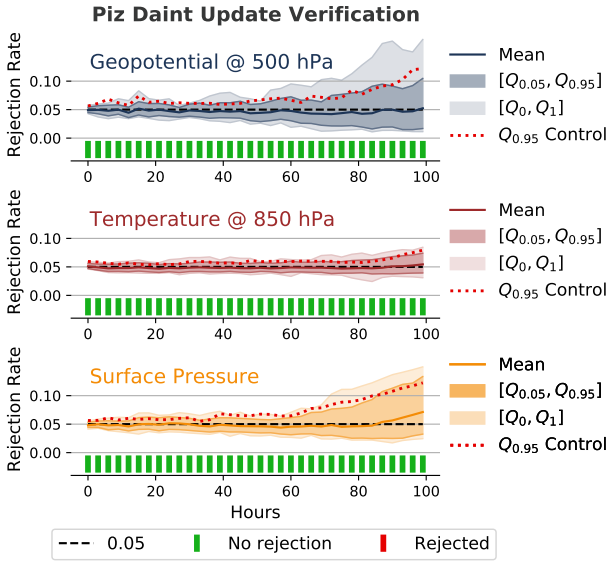

Figure 8. Rejection rates and decisions similar to Fig. 3 for the 500 hPa geopotential, 850 hPa temperature, and surface pressure from the verification of a major system update of the underlying supercomputer Piz Daint. The evaluation ensemble has been created with a model executable compiled and run after the update, whereas the verification and control ensemble have been produced by an executable compiled and run in an environment that replicates the state before the update. The methodology is clearly not able to reject the null hypothesis, which means that the update most probably did not significantly affect the model behavior.

in the diagnostic calculation of the field (see Sect. 4.3 for further detail). The focus on instantaneous values or averages over a small time frame (i.e. hourly for precipitation) is also a way to take internal variability into account. Very small differences can often only be detected during the first few hours or days, before the increasing internal variability outweighs the effect of the change. Therefore, we think that short simulations of a few days should be generally preferred to longer, computationally more expensive simulations.

It is not entirely clear how sensitive such a methodology is in the detection of differences in long climate simulations. For the verification of very slow processes, longer simulations with spatial and/or temporal averaging might appear to be the better choice. However, the current methodology using short integrations is also able to detect changes in slower variables such as soil moisture within the first few days, which indicates that it might also be suited for climate simulations. And given that differences from the frequent changes (e.g. compiler upgrades, library updates and minor code rearrangements) normally manifest themselves already early in the simulation (see Milroy et al., 2018), we think that this is a reasonable approach with low computational costs. Nevertheless, it is worthwhile to rethink our methodology in the case of a global coupled climate model that may represent very fast (e.g. the atmospheric model) and very slow (e.g. an ice sheet model) components. In such a case, it might be advantageous to test the different model components in stand-alone mode, before evaluating the fully coupled system.

https://doi.org/10.5194/gmd-2021-248
Preprint. Discussion started: 6 September 2021
© Author(s) 2021. CC BY 4.0 License.

Geoscientific Model Development Discussions

The choice of the tested variables in this work is quite arbitrary. Testing all possible model variables would guarantee the highest degree of reliability. But since the atmosphere is such a complex and interconnected system, most variables are highly correlated. Therefore, and based on our results, where we hardly see differences in test results for different variables, we think that testing a few standard output variables is already sufficient for a reliable result. Yet it is evident that carefully choosing several variables is essential. For instance, a model change in the formulation of groundwater runoff will take some time before it becomes detectable in the atmosphere. The incorporation of a principal component analysis and then applying the methodology only on the first few principal components (as in Baker et al., 2015) would also be a suitable option.

6 Conclusions and outlook

We presented an ensemble-based verification methodology based on statistical hypothesis testing that allows for an objective detection of model changes. The methodology operates on a grid-cell level and works for instantaneous and accumulated/averaged variables. With this high spatial and temporal resolution, the test results can already be used as a first informative basis for the debugging process. The study suggests that short-term ensemble simulations (days to months) are best suited, as the smallest changes are often only detectable during the first few hours of the simulation. Combined with the fact that the methodology already works well for coarse resolutions (here 50 km grid spacing), the methodology is a good candidate for a relatively inexpensive automated system test. We showed that the choice of the underlying statistical hypothesis test is secondary, as long as the rejection rate is compared to a rejection rate distribution from a control ensemble that has been generated with an identical statistical hypothesis test. While the methodology could theoretically be applied to all model output variables and thus be exhaustive, we think that this would be overkill. Based on our results using a limited-area climate model and the high correlations between many atmospheric variables, we think that a set of key variables such as 500 hPa geopotential, 850 hPa temperature, surface pressure, soil moisture, snow water equivalent, and precipitation might already be sufficient to cover most of the atmospheric and land-surface processes. For a fully-coupled global climate model, some further considerations will be needed.

The verification methodology was able to detect several configurations changes, ranging from very small changes, such as increased horizontal diffusion or changes in the minimum vertical heat diffusion coefficient, to bigger changes, such as disabling the subgrid-scale orography (SSO) parameterization. The test was not able to detect any differences between the regional weather and climate model COSMO running on GPUs or on CPUs on the same supercomputer (Piz Daint, CSCS, Switzerland). However, the test was able to detect differences between the double-precision version and the single-precision version of the model for 500 hPa geopotential and precipitation. In the case of 500 hPa geopotential, the rejection at the initial state suggests that there are already differences in the diagnostic calculation of the variable, whereas the rejection of precipitation is likely associated with differences in microphysics due to the reduced precision. Furthermore, the methodology has already been successfully applied for the verification of the regional weather and climate model COSMO after a major system update of the underlying supercomputer (Piz Daint, CSCS, Switzerland).

https://doi.org/10.5194/gmd-2021-248
Preprint. Discussion started: 6 September 2021
© Author(s) 2021. CC BY 4.0 License.

Geoscientific Model Development Discussions

Nonetheless, the results of such a test have to be interpreted with caution and might give a false sense of security. On the one hand, there is the issue associated with any statistical hypothesis test, where no rejection of the null hypothesis does not automatically mean that it is true. On the other hand, even though verification is termed as a "system test", it is hardly possible to test the whole model. There are usually countless different configurations for such models and testing all these configurations (i.e. different physical parameterizations, resolutions, numerical methods) is almost impossible and would require a huge computational effort. Similarly, applying the methodology to all possible model output fields would be computationally too expensive. The methodology also has some potential limitations in case a certain part of the code is only very rarely activated (as potentially the case with threshold-triggered processes).

For future work, we would like to apply the methodology for more test cases such as the compilation of the model with different optimization levels or running the model on different supercomputers. It would also be interesting to directly compare our verification methodology to other, already existing methodologies, in order to get a better idea about the differences in sensitivity and applicability (i.e. spatial and temporal scales).

Code and data availability. The source code that has been used to calculate the rejection rates shown in this paper is available under https://github.com/zemanc/verification_atmospheric_model. The corresponding model output data from the shorter ensemble simulations (10 days) is available under https://doi.org/10.5281/zenodo.5106467. The COSMO model that has been used in this study is available under license (see http://www.cosmo-model.org/content/consortium/licencing.htm for more information, last access: 16 July 2021). COSMO may be used for operational and for research applications by the members of the COSMO consortium. Moreover, within a liense agreement, the COSMO model may be used for operational and research applications by other national (hydro-)meteorological services, universities, and research institutes. ERA-Interim reanalysis data, which has been used for initial and lateral boundary conditions, is available at https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era-interim (last access: 16 July 2021).

Author contributions. CZ and CS conceptualized the verification methodology and designed the study. CZ performed the COSMO model ensemble simulations and developed the code for the verification of the model results. CZ wrote the paper with contributions from CS.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. We acknowledge PRACE for awarding compute resources for the COSMO simulations on Piz Daint at the Swiss National Supercomputing Centre (CSCS). We also acknowledge the Federal Office for Meteorology and Climatology MeteoSwiss, CSCS, and ETH Zurich for their contributions to the development of the GPU-accelerated version of COSMO.

https://doi.org/10.5194/gmd-2021-248
Preprint. Discussion started: 6 September 2021
© Author(s) 2021. CC BY 4.0 License.

References

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), Geoscientific Model Development, 8, 2829–2840, 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.

Baldauf, M., Seifert, A., Förstner, J., Majewski, D., Raschendorfer, M., and Reinhardt, T.: Operational Convective-Scale Numerical Weather Prediction with the COSMO Model: Description and Sensitivities, Monthly Weather Review, 139, 3887–3905, https://doi.org/10.1175/MWR-D-10-05013.1, 2011.

Bartlett, M. S.: The Effect of Non-Normality on the t Distribution, Mathematical Proceedings of the Cambridge Philosophical Society, 31, 223–231, https://doi.org/10.1017/S0305004100013311, 1935.

Bauer, P., Thorpe, A., and Brunet, G.: The quiet revolution of numerical weather prediction, Nature, 525, 47–55, https://doi.org/10.1038/nature14956, 2015.

Bellprat, O., Kotlarski, S., Lüthi, D., De Elía, R., Frigon, A., Laprise, R., and Schär, C.: Objective calibration of regional climate models: Application over Europe and North America, Journal of Climate, 29, 819–838, https://doi.org/10.1175/JCLI-D-15-0302.1, 2016.

Benjamini, Y. and Hochberg, Y.: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, Journal of the Royal Statistical Society: Series B (Methodological), 57, 289–300, https://doi.org/10.1111/j.2517-6161.1995.tb02031.x, 1995.

Clune, T. and Rood, R.: Software Testing and Verification in Climate Model Development, IEEE Software, 28, 49–55, https://doi.org/10.1109/MS.2011.117, 2011.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., Mcnally, A. P., Monge-Sanz, B. M., Morcrette, J. J., Park, B. K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J. N., and Vitart, F.: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system, Quarterly Journal of the Royal Meteorological Society, 137, 553–597, https://doi.org/10.1002/qj.828, 2011.

Doms, G. and Baldauf, M.: A Description of the Nonhydrostatic Regional COSMO-Model Part I : Dynamics and Numerics, https://doi.org/10.5676/DWD_pub/nwv/cosmo-doc_5.05_I, 2018.

Doms, G., Förstner, J., Heise, E., Herzog, H.-J., Mironov, D., Raschendorfer, M., Reinhardt, T., Ritter, B., Schrodin, R., Schulz, J.-P., and Vogel, G.: COSMO Documentation Part II : Physical Parameterization, https://doi.org/doi.org/10.5676/dwd_pub/nwv/cosmo-doc_5.05_ii, 2018.

Fuhrer, O., Osuna, C., Lapillonne, X., Gysi, T., Bianco, M., Arteaga, A., and Schulthess, T. C.: Towards a performance portable, architecture agnostic implementation strategy for weather and climate models, Supercomputing Frontiers and Innovations, 1, 44–61, https://doi.org/10.14529/jsfi140103, 2014.

Hong, S.-Y., Koo, M.-S., Jang, J., Kim, J.-E. E., Park, H., Joh, M.-S., Kang, J.-H., and Oh, T.-J.: An Evaluation of the Software System Dependency of a Global Atmospheric Model, Monthly Weather Review, 141, 4165–4172, https://doi.org/10.1175/MWR-D-12-00352.1, 2013.

https://doi.org/10.5194/gmd-2021-248
Preprint. Discussion started: 6 September 2021
© Author(s) 2021. CC BY 4.0 License.

Knight, C. G., Knight, S. H. E., Massey, N., Aina, T., Christensen, C., Frame, D. J., Kettleborough, J. A., Martin, A., Pascoe, S., Sanderson, B., Stainforth, D. A., and Allen, M. R.: Association of parameter, software, and hardware variation with large-scale behavior across 57,000 climate models, Proceedings of the National Academy of Sciences, 104, 12 259 LP – 12 264, https://doi.org/10.1073/pnas.0608144104, 2007.

Leutbecher, M. and Palmer, T. N.: Ensemble forecasting, Journal of Computational Physics, 227, 3515–3539, https://doi.org/https://doi.org/10.1016/j.jcp.2007.02.014, 2008.

Livezey, R. E.: Statistical Analysis of General Circulation Model Climate Simulation: Sensitivity and Prediction Experiments, Journal of Atmospheric Sciences, 42, 1139–1150, https://doi.org/10.1175/1520-0469(1985)042<1139:SAOGCM>2.0.CO;2, 1985.

Livezey, R. E. and Chen, W. Y.: Statistical Field Significance and its Determination by Monte Carlo Techniques, Monthly Weather Review, 111, 46–59, https://doi.org/10.1175/1520-0493(1983)111<0046:SFSAID>2.0.CO;2, 1983.

Lorenz, E. N.: Deterministic Nonperiodic Flow, Journal of Atmospheric Sciences, 20, 130–141, https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2, 1963.

Lott, F. and Miller, M. J.: A new subgrid-scale orographic drag parametrization: Its formulation and testing, Quarterly Journal of the Royal Meteorological Society, 123, 101–127, https://doi.org/10.1256/smsqj.53703, 1997.

Mahajan, S., Gaddis, A. L., Evans, K. J., and Norman, M. R.: Exploring an Ensemble-Based Approach to Atmospheric Climate Modeling and Testing at Scale, Procedia Computer Science, 108, 735–744, https://doi.org/https://doi.org/10.1016/j.procs.2017.05.259, 2017.

Mahajan, S., Evans, K. J., Kennedy, J. H., Xu, M., and Norman, M. R.: A Multivariate Approach to Ensure Statistical Reproducibility of Climate Model Simulations, in: Proceedings of the Platform for Advanced Scientific Computing Conference, PASC '19, Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/3324989.3325724, 2019.

Mann, H. B. and Whitney, D. R.: On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other, Ann. Math. Statist., 18, 50–60, https://doi.org/10.1214/aoms/1177730491, 1947.

Massonnet, F., Ménégoz, M., Acosta, M., Yepes-Arbós, X., Exarchou, E., and Doblas-Reyes, F. J.: Replicability of the EC-Earth3 Earth system model under a change in computing environment, Geoscientific Model Development, 13, 1165–1178, https://doi.org/10.5194/gmd-13-1165-2020, 2020.

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), Geoscientific Model Development, 11, 697–711, https://doi.org/10.5194/gmd-11-697-2018, 2018.

Pithan, F., Angevine, W., and Mauritsen, T.: Improving a global model from the boundary layer: Total turbulent energy and the neutral limit Prandtl number, Journal of Advances in Modeling Earth Systems, 7, 2029–2043, https://doi.org/10.1002/2015MS000503.Received, 2015.

Posten, H. O.: Robustness of the Two-Sample T-Test, in: Robustness of Statistical Methods and Nonparametric Statistics, edited by Rasch, D. and Tiku, M. L., pp. 92–99, Springer Netherlands, Dordrecht, https://doi.org/10.1007/978-94-009-6528-7_23, 1984.

Raschendorfer, M.: The new turbulence parameterization of LM, COSMO Newsletter, 1, 89–97, http://www.cosmo-model.org/content/model/documentation/newsLetters/newsLetter01/newsLetter_01.pdf, 2001.

Reichler, T. and Kim, J.: How Well Do Coupled Models Simulate Today's Climate?, Bulletin of the American Meteorological Society, 89, 303–312, https://doi.org/10.1175/BAMS-89-3-303, 2008.

Reinhardt, T. and Seifert, A.: A three-category ice scheme for LMK, COSMO Newsletter, 6, 115–120, 2006.

Ritter, B. and Geleyn, J.-F.: A Comprehensive Radiation Scheme for Numerical Weather Prediction Models with Potential Applications in Climate Simulations, https://doi.org/10.1175/1520-0493(1992)120<0303:ACRSFN>2.0.CO;2, 1992.

https://doi.org/10.5194/gmd-2021-248
Preprint. Discussion started: 6 September 2021
© Author(s) 2021. CC BY 4.0 License.

Rockel, B., Will, A., and Hense, A.: The regional climate model COSMO-CLM (CCLM), Meteorologische Zeitschrift, 17, 347–348, https://doi.org/10.1127/0941-2948/2008/0309, 2008.

Rosinski, J. M. and Williamson, D. L.: The Accumulation of Rounding Errors and Port Validation for Global Atmospheric Models, SIAM Journal on Scientific Computing, 18, 552–564, https://doi.org/10.1137/S1064827594275534, 1997.

Sandu, I., Beljaars, A., Bechtold, P., Mauritsen, T., and Balsamo, G.: Why is it so difficult to represent stably stratified conditions in numerical weather prediction (NWP) models?, Journal of Advances in Modeling Earth Systems, 5, 117–133, https://doi.org/10.1002/jame.20013, 2013.

Schär, C., Fuhrer, O., Arteaga, A., Ban, N., Charpilloz, C., Girolamo, S. D., Hentgen, L., Hoefler, T., Lapillonne, X., Leutwyler, D., Osterried, K., Panosetti, D., Rüdisühli, S., Schlemmer, L., Schulthess, T. C., Sprenger, M., Ubbiali, S., and Wernli, H.: Kilometer-Scale Climate Models, Bulletin of the American Meteorological Society, 101, E567–E587, https://doi.org/https://doi.org/10.1175/BAMS-D-18-0167.1, 2020.

Schättler, U., Doms, G., and Baldauf, M.: COSMO Documentation Part VII: User's Guide, https://doi.org/doi.org/10.5676/dwd_pub/nwv/cosmo-doc_5.05_vii, 2018.

Schlemmer, L., Schär, C., Lüthi, D., and Strebel, L.: A Groundwater and Runoff Formulation for Weather and Climate Models, Journal of Advances in Modeling Earth Systems, 10, 1809–1832, https://doi.org/10.1029/2017MS001260, 2018.

Storch, H. V.: A Remark on Chervin-Schneider's Algorithm to Test Significance of Climate Experiments with GCM's, Journal of Atmospheric Sciences, 39, 187–189, https://doi.org/10.1175/1520-0469(1982)039<0187:AROCSA>2.0.CO;2, 1982.

Student: The Probable Error of a Mean, Biometrika, 6, 1–25, https://doi.org/10.2307/2331554, 1908.

Sullivan, L. M. and D'Agostino, R. B.: Robustness of the t Test Applied to Data Distorted from Normality by Floor Effects, Journal of Dental Research, 71, 1938–1943, https://doi.org/10.1177/00220345920710121601, 1992.

Thomas, S. J., Hacker, J. P., Desgagn?, M., and Stull, R. B.: An Ensemble Analysis of Forecast Errors Related to Floating Point Performance, Weather and Forecasting, 17, 898–906, https://doi.org/10.1175/1520-0434(2002)017<0898:AEAOFE>2.0.CO;2, 2002.

Tiedtke, M.: A comprehensive mass flux scheme for cumulus parameterization in large-scale models, Monthly Weather Review, 117, 1779–1800, https://doi.org/10.1175/1520-0493(1989)117<1779:ACMFSF>2.0.CO;2, 1989.

Wan, H., Zhang, K., Rasch, P. J., Singh, B., Chen, X., and Edwards, J.: A new and inexpensive non-bit-for-bit solution reproducibility test based on time step convergence (TSC1.0), Geoscientific Model Development, 10, 537–552, https://doi.org/10.5194/gmd-10-537-2017, 2017.

Wicker, L. J. and Skamarock, W. C.: Time-Splitting Methods for Elastic Models Using Forward Time Schemes, Monthly Weather Review, 130, 2088–2097, https://doi.org/10.1175/1520-0493(2002)130<2088:TSMFEM>2.0.CO;2, 2002.

Wilcox, R. R.: Some practical reasons for reconsidering the Kolmogorov-Smirnov test, British Journal of Mathematical and Statistical Psychology, 50, 9–20, https://doi.org/https://doi.org/10.1111/j.2044-8317.1997.tb01098.x, 1997.

Wilks, D. S.: "The Stippling Shows Statistically Significant Grid Points": How Research Results are Routinely Overstated and Overinterpreted, and What to Do about It, Bulletin of the American Meteorological Society, 97, 2263–2273, https://doi.org/10.1175/BAMS-D-15-00267.1, 2016.

Zadra, A., Roch, M., Laroche, S., and Charron, M.: The subgrid-scale orographic blocking parametrization of the GEM Model, Atmosphere - Ocean, 41, 155–170, https://doi.org/10.3137/ao.410204, 2003.

Zeman, C., Wedi, N. P., Dueben, P. D., Ban, N., and Schär, C.: Model intercomparison of COSMO 5.0 and IFS 45r1 at kilometer-scale grid spacing, Geoscientific Model Development Discussions, 2021, 1–35, https://doi.org/10.5194/gmd-2021-31, 2021.

https://doi.org/10.5194/gmd-2021-248
Preprint. Discussion started: 6 September 2021
© Author(s) 2021. CC BY 4.0 License.

Zimmerman, D. W.: Comparative Power of Student T Test and Mann-Whitney U Test for Unequal Sample Sizes and Variances, The Journal of Experimental Education, 55, 171–174, https://doi.org/10.1080/00220973.1987.10806451, 1987.