Environmental Data Science (2023), 2: e22, 1–19
doi:10.1017/eds.2023.23

CAMBRIDGE UNIVERSITY PRESS

APPLICATION PAPER

Identifying climate models based on their daily output using machine learning

Lukas Brunner[1] and Sebastian Sippel[2,3]

[1]Department of Meteorology and Geophysics, University of Vienna, Vienna, Austria
[2]Institute for Meteorology, Leipzig University, Leipzig, Germany
[3]Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland
Corresponding author: Lukas Brunner; Email: l.brunner@univie.ac.at

Received: 17 May 2022; **Revised:** 14 March 2023; **Accepted:** 31 May 2023

Keywords: climate model evaluation; convolutional neural networks; logistic regression; machine learning; reanalysis

Abstract

Climate models are primary tools for investigating processes in the climate system, projecting future changes, and informing decision makers. The latest generation of models provides increasingly complex and realistic representations of the real climate system, while there is also growing awareness that not all models produce equally plausible or independent simulations. Therefore, many recent studies have investigated how models differ from observed climate and how model dependence affects model output similarity, typically drawing on climatological averages over several decades. Here, we show that temperature maps of individual days drawn from datasets never used in training can be robustly identified as "model" or "observation" using the CMIP6 model archive and four observational products. An important exception is a prototype storm-resolving simulation from ICON-Sapphire which cannot be unambiguously assigned to either category. These results highlight that persistent differences between simulated and observed climate emerge at short timescales already, but very high-resolution modeling efforts may be able to overcome some of these shortcomings. Moreover, temporally out-of-sample test days can be assigned their dataset name with up to 83% accuracy. Misclassifications occur mostly between models developed at the same institution, suggesting that effects of shared code, previously documented only for climatological timescales, already emerge at the level of individual days. Our results thus demonstrate that the use of machine learning classifiers, once trained, can overcome the need for several decades of data to evaluate a given model. This opens up new avenues to test model performance and independence on much shorter timescales.

Impact Statement

Climate models are used to study changes in the climate system and to inform decision makers. While models are getting better and better at representing the observed climate, some differences compared to observations remain. To isolate them from the influence of weather, such differences are typically investigated in climatological averages over several decades. Here, we show that maps from individual days are sufficient to robustly identify models using statistical and machine learning classifiers that have first been trained on different models and/or in a different time period. These results provide new ways of evaluating and interpreting model–observation and model–model differences, on short timescales.

1. Introduction

Multi-model ensembles, such as the latest sixth Coupled Model Intercomparison Project (CMIP6; Eyring et al., 2016), are widely used to investigate physical climate mechanisms, attribute past and project future changes, and inform political decisions. However, there is a growing awareness in the climate community that not all models included in CMIP provide equally plausible and independent simulations of the climate system (Tebaldi and Knutti, 2007; Knutti, 2010; Bishop and Abramowitz, 2013; Annan and Hargreaves, 2017; Eyring et al., 2019). Model evaluation methods, typically based on multi-decadal climatological averages to minimize the effect of internal variability on short timescales, have identified persistent biases across models compared to observations and highlighted the impact of model dependence on the similarity of model outputs (Masson and Knutti, 2011; Knutti et al., 2013; Boé, 2018; Bock et al., 2020; Brunner et al., 2020). Both model bias and similarities are ultimately due to the parametrizations used to represent processes that cannot be explicitly resolved and, to a lesser degree, due to the discretizations and numerical methods used to solve the fundamental model equations.

Work is ongoing to further reduce these parametrizations in new generations of kilometer-scale, storm-resolving global models with the ultimate aim of creating a digital twin of Earth (Bauer et al., 2021; Rackow et al., 2021; Hohenegger et al., 2022). These developments continue to blur the boundaries between weather prediction, climate prediction, and climate projection (Meehl et al., 2021) and call for new, innovative evaluation methods that allow models to be compared with observations and with each other on weather timescales. For example, biases in climate models can emerge over relatively short timescales when a model is initialized with an observed state, and pinpointing these biases may help model development (Palmer, 2016). This includes investigating the extent to which model biases and dependencies can be identified on such short timescales and how this can be interpreted.

A range of recent studies have demonstrated the potential of statistical and machine learning in climate science. These include approaches for seasonal prediction (Gibson et al., 2021), to identify forced signals from spatial patterns of temperature, precipitation, and humidity (Barnes et al., 2019; Sippel et al., 2020; de Vries et al., 2023), to explore the role of single forcing agents (Labe and Barnes, 2021), to predict modes of atmosphere–ocean variability (Gordon et al., 2021), and, most recently, to contrast models and observations (Labe and Barnes, 2022). Here, we investigate the potential of such statistical and machine learning classifiers to separate models from observations and from each other on the basis of daily data, and thus in the presence of considerable noise arising from the internal variability on weather timescales. We argue that the typically applied temporal aggregation over several years or even decades can be overcome and that future model evaluation methods might be able to be based on considerably shorter time periods, at least once classifiers have been trained. This can allow the investigation of new or updated models for which decades of data are not yet available, as showcased by the inclusion of 1 year of experimental storm-resolving simulations provided by the NextGEMS project.[1]

The work is guided by two main questions which are detailed and discussed in Sections 3 and 4: based on daily temperature output (a) can out-of-sample test datasets (i.e., datasets never used in training) of models and observations reliably be identified as different from each other and (b) can temporally out-of-sample test data be identified by their name, even if they come from a different climate regime? The datasets and methods are presented in Section 2, and the results are summarized and interpreted in Section 5.

2. Data and Methods

2.1. Model and observational data

We use all available CMIP6 models which provide daily data for 2 m surface air temperature in the historical period. In total, this amounts to 43 models, which can be broadly grouped into 22 families by developing institutions (see Supplementary Table S3 for a full list). To represent the observations, we use

[1] https://nextgems-h2020.eu

https://doi.org/10.1017/eds.2023.23 Published online by Cambridge University Press

four datasets selected to cover some of the diversity in observational products: (a) the fifth generation of the European Centre for Medium-Range Weather Forecasts (ECMWF) Retrospective Analysis (ERA5; Hersbach et al., 2020); (b) the Modern-Era Retrospective analysis for Research and Applications version 2 (MERRA2; GMAO, 2015; Gelaro et al., 2017); (c) the Twentieth Century Reanalysis version 3 (20CR; Slivinski et al., 2021), which assimilates only surface pressure observations and, hence, relies considerably more on modeling for its output; (d) the Daily Optimum Interpolation Sea Surface Temperature dataset version 2 (DOISST; Huang et al., 2021), which could be considered as more direct observations compared to the other datasets. The first three observational datasets provide 2 m surface air temperature also over the oceans equivalent to the models, while DOISST is based on interpolated in situ and remote sensing measurements of sea surface temperature, which is a slightly different but closely related measure of temperature.

The time period from 1982 to 2014 is used to match the first availability of all observational datasets and the end of the historical forcing period in CMIP6, respectively, and all daily temperature fields (including the high-resolution simulations) are regridded to 2.5° × 2.5° resulting in a total of 10,368 grid cells (72 latitudes × 144 longitudes). Since DOISST does not provide data on land, we apply a common land-sea mask to all datasets, using only the 6,888 ocean grid cells. Sensitivity analysis (omitting DOISST) shows that results are similar if the analysis is based on all grid cells.

In addition, we use temperature projections for the end of the century (2091–2100) driven by the high emission scenario SSP5-8.5 (Meinshausen et al., 2020), where available, to test the robustness of our approach under severe warming (see Supplementary Table S3). Finally, we also draw on 1 year (February 2020 to January 2021) of prototype data from cycle one of the NextGEMS project and use global temperature fields from an ICON-Sapphire run with an atmospheric resolution of 5 km (experiment ID: dpp066; Hohenegger et al., 2022).

2.2. Statistical and machine learning classifiers

We use two different statistical and machine learning methods to separate models from observations and from each other. First, logistic regression, which allows insights into the learned coefficients but has the limitation of being a linear method. Second, a convolutional neural network (CNN) which represents rather the other end of the complexity spectrum, being able to learn nonlinear spatial relations between features but lacking the easy interpretability of logistic regression.

Logistic regression is linear in its parameters and takes an $M \times N$ matrix as input, where N is the number of samples (days) and M is the number of features (ocean grid cells). For training an additional vector of length N is provided containing the true labels y_n (i.e., 0 for "model" and 1 for "observation") for each sample. A continuous predicted probability $\widehat{p}_n$ is then assigned to each sample X_n based on its features $x_{n,m}$ and the regression parameters w_m via the logistic function:

$$\widehat{p}_n = \frac{1}{1+\exp\left(-\left(w_0+\sum_{m=1}^{M} w_m x_{n,m}\right)\right)}$$

The outcome is in $[0,1]$ with 1 indicating the highest chance that a given day belongs to the "observation" category and at the same time the lowest chance that belongs to the "model" category since the categories are complementary:

$$\widehat{p}(y_n = 0|X_n) = 1 - \widehat{p}(y_n = 1|X_n)$$

The predicted category $\widehat{y}_n$ is based on this probability $\widehat{p}_n$ with a decision threshold of 0.5. The logistic regression classifier is constrained by a L_2 regularization (i.e., a penalty on the squared sum of regression coefficients w_m) to avoid overfitting and to ensure smooth coefficients in space. Overall coefficients w_m are searched to minimize the expression:

$$\min_w \left\{ -C \sum_{n=1}^{N} [y_n \log(\widehat{p}_n) + (1-y_n)\log(1-\widehat{p}_n)] + \sum_{m=0}^{M} w_m^2 \right\}$$

https://doi.org/10.1017/eds.2023.23 Published online by Cambridge University Press

Note that in the implementation based on scikit-learn[2] used in this paper, the regularization parameter C is multiplied with the residuals rather than with the coefficients, meaning that smaller values of C lead to stronger regularization. C is optimized using 5-fold cross-validation. More general details on regularized logistic regression can be found, for instance, in Hastie et al. (2009).

To complement the logistic regression classifier, we use a CNN as a second method. Deep neural networks, such as CNNs, can have considerably more trainable parameters (often organized in multiple layers) and can be less interpretable than traditional methods. This means that they can be, on the one hand, more prone to overfitting but, one the other hand, can also learn more complex, nonlinear relationships in the data. Therefore, their use in different scientific disciplines, not least in climate sciences, has been rapidly increasing in recent years (Kashinath et al., 2021; Hsieh, 2022). Due to their complexity, many different design choices in the exact layout of the network are possible; here, we use an out-of-the-box setup for image classification without hyperparameter tuning but adjusted to the resolution of the daily temperature maps used in the input layer. Overall, the CNN consists of an input layer, eight hidden layers, and an output layer. See Supplementary Table S2 for details about the layout.

The 2-dimensional temperature fields from each day can directly be interpreted by the CNN equivalent to a image classification task; therefore, the input layer takes a $K \times L \times N \times 1$ matrix where N is, again, the number of days, K is the number of latitudes, L is the number of longitudes, and 1 is a single color channel representing the temperature values. Note that for this case land grid cells are included but set to a constant fill value. The convolutional layers use the rectified linear unit as activation function, while the output layer uses the softmax operator to assign probabilities to each of classification categories z_n:

$$\widehat{p}_n = \frac{e^{z_n}}{\sum_n^N e^{z_n}}$$

This means that, in contrast to the logistic regression, the CNN is set up to separate also between multiple output categories (multi-class case) and a given input is identified as belonging to the class with the highest probability. For the binary case with only the "model" and "observation" classes, discussed in the first part of this paper, this means that the probabilities are complementary and the threshold for assignment is 0.5 as for the logistic regression. To avoid overfitting for the CNN, part of the training data is used for validation during the training process and an early stopping criterion is applied on the validation loss. The evolution of loss and accuracy during the training epochs are shown in Supplementary Figures S2 and S3. All classifiers are mostly well-calibrated with the CNN showing a tendency for overconfidence, in particular, for the multi-class cases (Supplementary Figures S4–S6).

The performance of the classifiers is, on the one hand, assessed using the overall accuracy which is defined as the number of correct predictions divided by the number of total predictions. On the other hand, also the confidence in the predictions is assessed based on the probabilities assigned to each test sample from each dataset individually.

2.3. Out-of-sample frameworks and preprocessing

We use daily, land-masked temperature fields as samples, resulting in 6,888 grid cells being used as features for the classification. The 2-dimensional daily fields are either used directly in the case of the CNN classifier or flattened to a 1-dimensional feature array in the case of logistic regression. Note that the changing grid cell area with latitude due to longitude convergence is not explicitly accounted for in this study but may be implicitly learned by the classifiers.

Training and validation days are drawn from the 20-year period 1982–2001, as test data all days from the two *temporally out-of-sample* 10-year periods 2005–2014 and 2091–2100 are used. Therefore, all results are based on test samples that were never seen in training.

[2] https://scikit-learn.org

https://doi.org/10.1017/eds.2023.23 Published online by Cambridge University Press

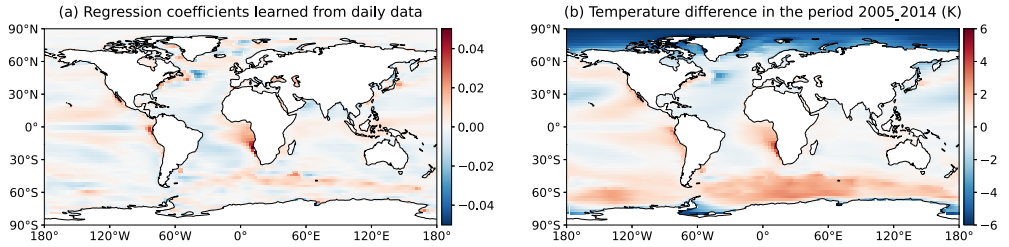

Figure 1. *(a) Logistic regression coefficients learned from 17,200 randomly drawn daily samples in the period 1982–2001 to separate models and observations. (b) Climatological mean, multi-model mean temperature difference to the mean over the four observational datasets in the period 2005–2014. See Supplementary Figure S7 for corresponding maps of the individual models. Coefficients and climatologies are calculated from daily data with the global mean removed.*

For the separation of "model" versus "observation" in Section 3, we use an additional *dataset out-of-sample* framework to explore whether the learned parameters can be generalized to unseen datasets: a separate classifier is trained for each dataset, where all samples from the dataset in question are withheld from training. To be even stricter, for climate models not only the model in question is withheld in training, but also all closely related models (see Supplementary Table S3 for a list of the model groups). We chose this approach because it has been shown that closely related models (i.e., models developed at the same institution) can have very similar output (Brunner et al., 2020). The only exceptions to this are the regression coefficients shown in Figure 1 and the bootstrap tests for which all datasets were used in training.

In both machine learning and climate modeling, there are different preprocessing and bias-correction options and corresponding terminologies. These include more general approaches such as feature normalization and more domain-specific approaches such as bias correction. The latter are based on physical system understanding, such as the insight that the mean temperature bias is not a relevant predictor of a model's ability to simulate changes in the climate system (Giorgi and Coppola, 2010). Here, we opt for two domain-specific bias-correction approaches from climate sciences: (a) from each daily temperature field, the global mean over that field is removed and (b) from each daily temperature field, the seasonal mean field is removed in addition to (a).

For (b), the seasonal mean temperature field is calculated individually for each dataset and only based on the training data. For each day-of-the-year and grid cell, it is calculated as the average over ±15 days around that day-of-the-year and over all training years (i.e., a mean over 31 days × 20 years = 620 values). For the ICON-Sapphire model, the seasonal cycle is estimated based on only 1 year (i.e., 31 values centered around each day of the year).

Finally, each sample is associated with one of two label types to be predicted by the machine learning classifiers: either "model"/"observation" (binary case) or the name of the dataset (multi-class case; see the dataset ID column in Supplementary Table S3). A summary of all classifiers, bias corrections, and out-of-sample strategies employed in this work can be found in Supplementary Table S1. The living code that implements the general method is available on GitHub (https://github.com/lukasbrunner/model_learning) and from Brunner and Sippel (2023). Supplementary Figures S7 and S8 show an example test day from each of the datasets and for the two preprocessing cases, respectively. While, in the first case, latitudinal temperature gradients due to differing solar insolation clearly dominate the temperature pattern, in the second case patches of regionally cooler or warmer temperatures emerge, related to the synoptic atmospheric conditions on that day. Note that on a given day (March 21st 2010 in Supplementary Figures S7 and S8) models are not expected to simulate identical weather patterns compared to the observations. The models are free-running and, therefore, simulate different synoptic situations and related temperature patterns while the four observation-based datasets assimilate measurements from the same day leading to very similar surface temperature patterns.

https://doi.org/10.1017/eds.2023.23 Published online by Cambridge University Press

3. Separating Models and Observations

3.1. Regression coefficient maps to separate models and observations

First, we train a single logistic regression classifier on all available datasets to establish the regularization parameter and examine the learned coefficients. For this case, we use data with only the global mean removed. We use 200 different, randomly drawn training days from each of the 43 models resulting in 8,600 training samples labeled "model" which are matched by 8,600 random days labeled "observation" (2,150 from each of the four observational datasets). The classifier manages to correctly identify the vast majority of test samples for this setup so that the number of training samples was limited to this amount to save computational resources. It can be assumed that increasing the number of training samples would slowly improve the classification skill even further. The results are also robust to using different random training samples (see Section S3 in the Supplementary material for results from a 100-member bootstrap test).

Since temperatures between neighboring grid cells, used as features, are not independent, the 5-fold cross-validation yields a strong L_2 regularization parameter of about $C = 0.002$, which is used for all logistic regression classifiers in this section. The spatial patterns in the features that are important for separating models and observations are reflected in the regression coefficients learned by the classifier and are shown in Figure 1a. The distribution of the coefficients identifies areas important for the separation of climate models and observations on a daily scale. In addition, the sign of the coefficients can be interpreted physically, with positive values indicating regions where models tend to be warmer than the observations and vice-versa.

The most prominent region of negative coefficients is found in the North Atlantic, near the so-called North Atlantic warming hole (Chemke et al., 2020; Keil et al., 2020). Here models appear to systematically underestimate temperatures on a daily basis compared with observations and relative to the global mean. In contrast, there are regions of high coefficients at the eastern edges of the Pacific and Atlantic ocean basins. These regions correspond to persistent model biases in the representation of clouds and their radiative effects which are known to occur on all timescales from daily to decadal (Williams et al., 2013; Hsi Yen Ma et al., 2014; Brient et al., 2019; Bock et al., 2020; Chen et al., 2022). In the equatorial Pacific, known as a region with notorious climate model biases that typically show too cold and too narrow equatorial cold tongues, negative coefficients are also found by the logistic regression classifier. This is accompanied by warm biases to the north and south (shown as positive regression coefficients) associated with the models' representation of the intertropical convergence zone (Hirota et al., 2011; Li and Xie, 2014; Tian and Dong, 2020). Overall, there are notable consistencies in several of the large-scale patterns between the logistic regression coefficients and the climatological mean, multi-model mean biases (Figure 1a,b).

However, there are also several regions where the two do not match, such as high northern latitudes, parts of the southern ocean, and the Antarctic coast. This can be an indication that the corresponding features are not shared across all (or at least most) potential days drawn from different models and from across the seasonal cycle. The high northern latitudes are briefly discussed here as one example for such a case with the clear climatological cold bias in the multi-model mean (Figure 1b) not being reflected in corresponding patterns in the regression coefficients (Figure 1a). This is probably caused by a combination of reasons as this region is known for its large (climatological) model spread (see, e.g., Notz and Community, 2020 and Supplementary Figure S9), its seasonally varying biases with cold biases mainly coming from the winter season (Davy and Outten, 2020), and large internal variability which reduces the size of the regression coefficients there (Barnes et al., 2019; Sippel et al., 2020). In general, it is, therefore, not a priori clear if patterns important for the classification of individual days and long-term climatological biases (averaged over multiple models) would match at all, but our findings show that for several regions this is the case.

3.2. Logistic regression classification of out-of-sample datasets

In the rest of this section, we use the *dataset out-of-sample* approach described in Section 2.3 to show whether the classifiers can be generalized to datasets unseen in training. The probabilities assigned to the

https://doi.org/10.1017/eds.2023.23 Published online by Cambridge University Press

test samples by each corresponding classifier are aggregated in Figure 2a. For the vast majority of cases, the logistic regression classifiers assign the correct category with close to 100% probability leading to an overall accuracy of 99.4% (excluding ICON-Sapphire discussed below). Some of this skill may be due to remaining dependencies across families, which were quite loosely defined based on institutions in this study. However, it could also be an indication for the existence of persistent distinguishing features that can be transferred between models even in presence of the large internal variability on daily timescales.

For several models, a fraction of test samples is predicted with less certainty (boxes and whiskers emerging from the zero-line in Figure 2a and reliability diagram in Supplementary Figure S4). The model families most prone to get confused with observations are CMCC, CNRM, EC-Earth3, GFDL, and INM. An intuitive interpretation of this behavior might be that these models provide the "best" representation of "true" temperatures on a daily basis as they are most similar to the observations (at least viewed through the lens of logistic regression). However, this is not conclusive and would require additional evidence. First, there are only very few samples with a nonzero probability of belonging into the observation category for any given model, so the confusion could be due to chance (i.e., a model sample could be classified as observations, but for the wrong reasons so that no conclusion should be drawn about the overall performance of the model). A second important consideration is possible codebase overlap also between models and the observational datasets, which could be picked up by the classifier as discussed in more detail in Section 4. Nevertheless, the approach of intentionally exploiting misclassifications of models as observations is still a promising avenue for future research as recently also highlighted by work from Labe and Barnes (2022), who find that for the Arctic test samples observations are most prone to get confused for certain models.

Focusing on model families that provide high and low resolution variants of the same model (AWI, CMCC, CNRM, HadGEM, MPI, and NorESM2), one can speculate about some resolution dependence in the probability to be misclassified. Several finer-resolution model variants appear to have a higher chance to be misclassified compared to their coarser-resolution siblings (model names including HR, MR/MM, or LR/LL indicating relatively higher to lower resolution in Figure 2a). Such a behavior would be

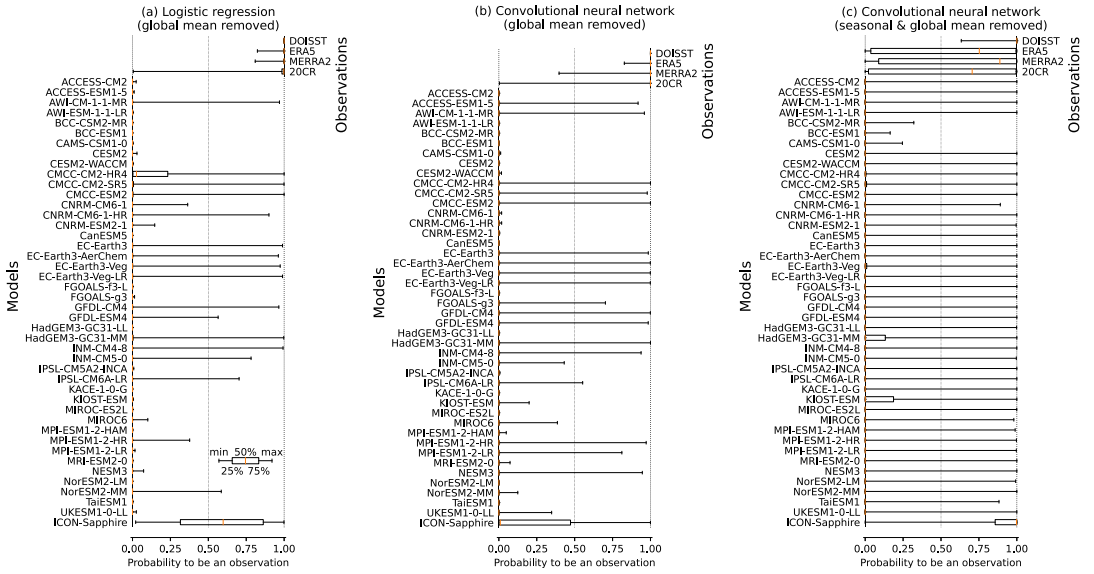

Figure 2. *Distribution of predicted probabilities for the dataset out-of-sample test days: for each dataset, the probabilities are estimated by a classifier which has not been trained on this dataset. The vertical dotted line at 0.5 marks the decision threshold between the two categories. ICON-Sapphire is never used in training and has only 1 year of data available. (a) Results for logistic regression classifiers using data with the daily global mean removed. (b) Same as (a) but for the convolutional neural network. (c) Same as (b) but using data with the seasonal cycle removed in addition.*

https://doi.org/10.1017/eds.2023.23 Published online by Cambridge University Press

consistent with studies based on climatological timescales such as the findings of Bock et al. (2020), who note that long-standing regional model biases are smaller in higher resolution versions of the same model.

3.3. Classifying samples from the kilometer-scale ICON-Sapphire model

To investigate this further and to highlight a potential application of our approach, we include preliminary results from the NextGEMS project and predict samples from 1 year of data from a global, storm-resolving (atmospheric resolution 5 km) simulation using ICON-Sapphire (Hohenegger et al., 2022). Due to the high resolution, processes that need to be parameterized in CMIP6-type models can be explicitly resolved in ICON-Sapphire, which can be expected to have considerable impacts also on the daily temperature fields. However, they do not obviously show in the climatological mean difference to the observations which is comparable to coarser resolved models and shows similar patterns, although we note that this could be a coincidence given that only a single year is used (see Supplementary Figure S10). This means that based on the comparison of climatologies alone, one might assume that the logistic regression classifier should be able to unambiguously identify ICON-Sapphire as a model.

To investigate this, we classify test samples from ICON-Sapphire using the logistic regression classifier trained on all other datasets. However, despite the similar climatologies, the classifier is unable to clearly identify ICON-Sapphire as either model or observation with about half of the samples predicted to be in either category. This indicates that the classification is (at least partly) based on more complex relationships in the high-dimensional feature space than can be easily assessed by the comparison shown in Figure 1. The explicit resolution of processes which are parameterized in coarser models seems to lead to differences in the daily temperature fields in ICON-Sapphire that prevent a correct classification, even though they do not clearly emerge in the climatological mean. This points to a potentially highly encouraging emergent behavior of the ICON-Sapphire simulations, but in principle some effect of compensating errors cannot be excluded, and hence we warrant a careful interpretation of this result, which will be verified as soon as a longer simulation with this model is available. Should such a relationship hold in future research, it will enable innovative, new ways of model evaluation based on much shorter timescales than the 20+ years typically used.

To test whether the results for the ICON-Sapphire model are merely an artifact of the logistic regression, we also use a more complex, but less interpretable CNN. The CNN classifiers achieve a similar overall accuracy of 99.7% for the same *dataset out-of-sample* framework and samples with the global mean removed. In general, the pattern of models with samples that get misclassified is quite consistent between the logistic regression and the CNN (Figure 2a,b). While ICON-Sapphire still stands out as the model most likely to be confused for an observation, confirming the results from the logistic regression classifier, the CNN is able to correctly identify the majority (75.1%) of test samples, suggesting that it learns some more fundamental model properties that persist also at high resolution. Compared to the rather simple "model fingerprint" shown for the logistic regression in Figure 1a, the classifications from the CNN are, thus, likely to be based on more complex relationships. We plan to investigate these relationships and the importance of different grid cells for the skill of the CNN in future work, drawing on techniques that, for example, aim to reveal regions of higher and lower importance in the temperature maps used as input (Bach et al., 2015).

3.4. CNN classification of out-of-sample datasets without climatological bias

Based on these results, we, next, test if models and observations can still be separated in the absence of any climatological biases. To do this, we now also remove the mean seasonal cycle from each sample (see Section 2.3 for methodological details, Supplementary Figure S11 for the resulting multi-model mean bias equivalent to Figure 1b, and Supplementary Figure S12 for a breakdown by individual models). This means that any dataset-specific persistent regional biases that might have served as a basis for separation so far are now removed, along with any biases in the equator-pole gradient or between the hemispheres. Therefore, the classifiers can now only train on the spatial relationships of the remaining internal

https://doi.org/10.1017/eds.2023.23 Published online by Cambridge University Press

variability (which could be interpreted as daily weather), making the classification task considerably harder.

For this case, logistic regression no longer has any skill as the only remaining sources of information are nonlinear relations between the spatial structures of daily global weather and the test samples are all centered around the decision threshold (not shown). In contrast, the CNN achieves an overall accuracy of about 94.2% demonstrating the power of this nonlinear method. Figure 2c shows the corresponding breakdown of predicted probabilities revealing that now almost all models get confused for observations a number of times but still all are classified correctly for the vast majority of test samples, with the sole exception being the ICON-Sapphire model.

For ICON-Sapphire, the seasonal cycle has been estimated using only the 31-day running window as only 1 year is available. This differs from the other datasets where the seasonal cycle has been calculated over the full 20 years of the training period (Section 2.3). Therefore, the results for ICON-Sapphire need to be interpreted with care and should be revisited once more data are available. Nevertheless, we show these preliminary findings here, to highlight that, based on these results the structure of the remaining internal variability in ICON-Sapphire is recognized as more closely resembling observations than CMIP6 generation models. In fact, ICON-Sapphire is more frequently classified as observation than three of the four observational datasets (ERA5, MERRA2, and 20CR) by the respective classifiers trained with the tested dataset withheld. While this result is preliminary and not conclusive due to the limited amount of data available, it, again, points toward very encouraging properties of kilometer-scale models that warrant closer investigation as more data become available.

In turn, the DOISST dataset is identified correctly with perfect accuracy in all three cases shown in Figure 2 with only individual misclassifications appearing even when bootstrapping the training data (Supplementary Figure S1). Optimistically interpreted, this could mean that the classifiers are picking up on the fact that DOISST is the dataset with the least amount of model included (see Section 2.1) and that they have indeed learned some fundamental distinguishing features. This interpretation is supported by the fact that 20CR is most prone to get confused for a model, while also being the observational dataset that relies most heavily on a model for its output.

The overall skill of this binary classification of out-of-sample datasets is notable, in particular, when considering that bias correcting each model by subtracting the mean seasonal cycle and the daily global mean effectively removes the entire time-persistent regional bias as well as any global mean offset between the datasets. This means that the only remaining sources of information to learn from are amplitude and spatial dependencies of the remaining daily temperature variability. One remaining source of model-observation differences for this case could arise from the coupling of atmosphere and ocean and the resulting surface energy balance in the models.

In a planned follow-up study, we will build on the results presented here and analyze the origin of this skill in more detail, using explainable machine learning techniques (e.g., Bach et al., 2015) as well as more specific approaches drawing on domain knowledge from climate sciences. The latter will include, for example, targeted masking of certain regions where climate models are known to perform particularly well/poorly to systematically investigate which areas of the globe are essential for skill and how this might depend on the models used. Such a combination of general and domain-specific approaches to classification skill is important to account for special properties of the temperature maps used compared to more general image classification. These include fundamental properties of the climate system such as the imprint of topography, temperature gradients, and circulation patterns, which are to some extent common to all datasets.

4. Identifying Models by Name

In the first part of this paper, we showed that there are common features across models and observations, respectively, that enable us to reliably identify even datasets unseen in training. In this section, we investigate whether there are also separating features that allow us to distinguish models from each other. Previous research, based on climatological timescales, has shown that models can be separated as well as clustered into families based only on their output (e.g., Masson and Knutti, 2011; Knutti et al.,

https://doi.org/10.1017/eds.2023.23 Published online by Cambridge University Press

2013; Boé, 2018; Brunner et al., 2020; Merrifield et al., 2020, 2023). Here, we investigate whether models (and observations) still have unique features that allow them to be identified even on daily timescales.

We use the CNN on the data with the seasonal cycle and global mean removed and train it to recognize each of the 43 models as well as the four observational datasets. We increase the number of training samples to 2,000 per dataset for this case where we only train a single classifier. In the previous section, the tested dataset was withheld from training, this is obviously no longer possible for this case where we aim to identify each model by name. Therefore, we only use temporally out-of-sample test samples (see Section 2.3 for details).

Assigning the correct label to each of the 47 datasets yields an overall accuracy of 83.4%. To put this into context, note that compared to the binary classification in the last section, we are now aiming to separate 47 categories, which considerably increases the difficulty of the classification. The CNN is thus able to pick up patterns unique to each dataset in order to separate it from all other datasets.

Figure 3 shows a breakdown of accuracy by dataset in a confusion matrix of true sample names versus the predicted sample names. Correct predictions (predicted name equals true name) are located on the main diagonal of the matrix and use green shading. As expected from the overall accuracy, the majority of samples are assigned correctly with misclassifications exceeding 10% almost exclusively only found within model families. Such misclassifications between models from the same family are shown in purple shading and are mostly located in the secondary diagonals, as most related models have similar names and the models are ordered alphabetically.

Overall, the model with the highest number of correctly identified samples is MIROC-ES2L (99.9%), indicating that it is very different from all other models. This is consistent with studies using time averages over several decades to investigate model dependence in CMIP5 (Knutti et al., 2013) and CMIP6 (Brunner et al., 2020). In turn, the models with the lowest number of correct predictions all belong to larger model families (models with less than 80% accuracy in Figure 3). These results highlight that models can be separated by name on the basis of their pattern of daily internal variability and that overlaps in the models' source code lead to similarities in their daily internal variability, resulting in a higher chance for misclassification within families.

On closer inspection, even some of the misclassifications outside of model families (red shading in Figure 3) follow (more distant) model relationships. For example, about 10% of samples from the UK's HadGEM3-GC31-LL are misclassified as the Australian ACCESS-CM2 model (middle of first column in Figure 3), which is probably due to the fact that ACCESS-CM2 reuses many of the UK models' components (Bi et al., 2020). Similarly, about 7% of samples from ACCESS-CM2 are misclassified as the Korean KACE-1-0-G model (middle of first row in Figure 3), which is also related to the HadGEM family (Lee et al., 2020). Similar considerations apply to other related model groups (see, e.g., Brunner et al., 2020 for a discussion of broader model families), although there also remain a number of misclassifications that cannot be explained.

Similar to the binary case discussed in the first part, resolution (and related changes in the parametrizations) emerges as a property that is important for classification skill. For example, in the CMCC family the CM2-SR5 and ESM2 variants are confused with each other in about a third of the samples while the HR4 variant with a higher resolved ocean (Cherchi et al., 2019) is hardly confused with either of the former two, suggesting that the higher resolution sets the model apart. A somewhat similar pattern can be observed for the other three families with three or more members: EC-Earth (Veg-LR is less often confused with family members and has a lower atmospheric resolution; Döscher et al., 2022), HadGEM (GC31-MM with higher atmospheric and ocean resolution; Andrews et al., 2019), and MPI (HR with higher atmospheric and ocean resolution; Mauritsen et al., 2019).

Concerning observations, a notable feature of Figure 3 is that misclassifications between models and observations are not symmetric. Hardly any models are being misclassified for observations, while the observational datasets ERA5 and MERRA2 get confused as models for 20 and 30% of samples, respectively. This is consistent with the dataset out-of-sample results shown in Figure 2c, where we found that observations are mistaken for models more often than vice-versa. A possible interpretation of

https://doi.org/10.1017/eds.2023.23 Published online by Cambridge University Press

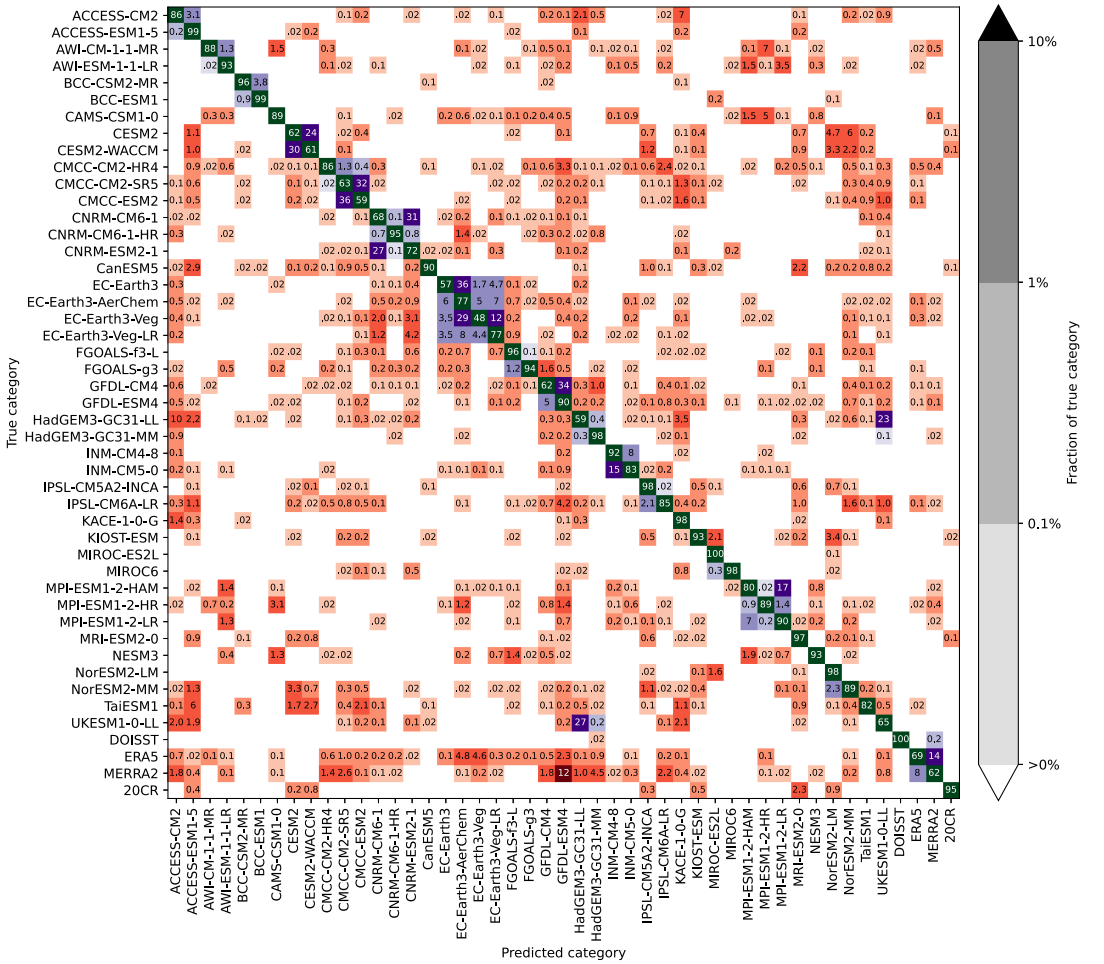

Figure 3. *Confusion matrix showing the frequency of predicted versus true labels. The main diagonal shows correct predictions using green shading, purple shading indicates misclassifications within a model family (see Supplementary Table S3), and red shading indicates other misclassifications. Values are in % relative to the total number of samples in each category. The number in each box gives the value rounded to the last shown digit with rows not adding up to 100% only due to rounding.*

this behavior is that models produce more homogeneous patterns persistent across days (samples), while the two reanalyses in question produce more diverse output due to the assimilation of observations. This will be further investigated in future work, including the identification of possible patterns in the misclassifications, for example, a seasonality.

Similar to confusions between models, observational misclassifications may reflect potential (remote) dependencies in the source code even between models and reanalyses. For example, ERA5, for which about 10 % of test samples are predicted to be from the EC-Earth family, which has documented dependencies on the ECMWF atmosphere which is also used in ERA5. (Döscher et al., 2022). MERRA2 is predicted to be the GFDL-ESM4 model for about 12% of cases, which might be attributable to common heritage as the atmospheric models used in GFDL-ESM4 (AM4.0; Dunne et al., 2020) and MERRA2 (GEOS-5; Rienecker et al., 2008; Molod et al., 2015) are based on the same dynamical core (Lin, 2004).

https://doi.org/10.1017/eds.2023.23 Published online by Cambridge University Press

The third observational dataset, DOISST, is never misclassified which is consistent with the results from the binary case presented in Figure 2. The combination of these two results suggests that DOISST has very clear observational properties but is still a very distinct dataset. For 20CR, we find a slightly different behavior to the binary case, with the number of misclassifications being considerably reduced in Figure 4. This could mean that 20CR has properties of both model and observation, making it easier to identify it as an individual dataset rather than as belonging into the broader observational category.

Finally, we investigate how the daily differences between models relate to the differences due to global warming and check whether the classification remains robust in a changing climate by drawing test samples from the period 2091–2100 rather than from 2005 to 2014. There is no absolute warming present in either training or test samples due to the removal of the daily global mean, but daily weather patterns are expected to change significantly in a warming world (e.g., Sippel et al., 2020). This is particularly true as we use data from the high-emission pathway SSP5-8.5 which leads to an additional global mean warming of about 4 K compared to today depending on the model (IPCC, 2021). For this case, only 33 models and no observations are available. Figure 4 shows the patterns learned by the CNN in the historical period still persist even after severe climate change and allow the correct identification of about 69.9% of the test samples from the end of the century. For this case, several models have more than 10% of their samples

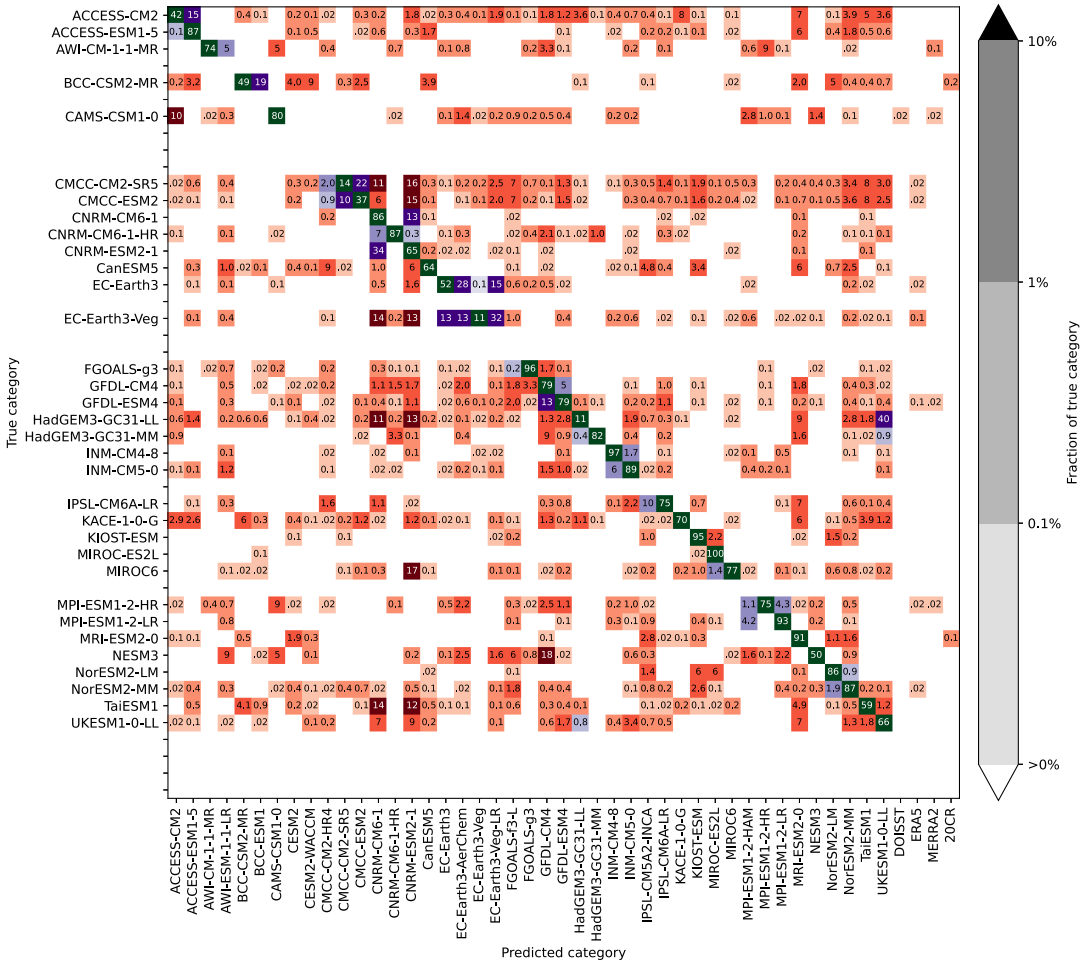

Figure 4. *Same as Figure 3 but with test data from the end of the century (2091–2100). Labels from datasets which do not cover this period are omitted in the true category.*

https://doi.org/10.1017/eds.2023.23 Published online by Cambridge University Press

misclassified as another, not closely related model. Still, many of the characteristics discussed above persist for this case. For example, MIROC-ES2L remains the model with the highest accuracy and EC-Earth3-Veg remains the model with the most misclassified samples.

Ultimately, classifiers similar to the one used here, once trained, could be used, for example, in model development to investigate the impact of changes in parametrizations, resolution, or model components on the model output without the need to run the tested model for several decades. Conversely, if a model is no longer recognized by name in a future run, this might also be an indication for a very different behavior and spatial patterns, thus warranting a closer investigation.

5. Summary, Conclusions, and Outlook

We have shown that machine learning classifiers, once trained, can distinguish CMIP6 generation climate models from observations based on only the temperature map of a single day even for entirely new models never used in training. Both investigated approaches (logistic regression and CNN) were found to perform very well in separating 2 m surface air temperature fields where only the global mean has been removed, both achieving an accuracy of over 99%. When the mean seasonal cycle was removed from each grid cell and dataset in addition, the logistic regression no longer showed any skill while the CNN still achieved an accuracy of over 90%.

The performance for the latter case is quite remarkable, given that the removal of the mean seasonal cycle means that the classifiers could not learn from regions of dataset-specific, time-persistent biases, and thus had to rely only on the spatial dependence structures and the amplitudes of variations of daily global weather. The properties of the CNN as a highly nonlinear, deep learning method, however, did not allow a straightforward extraction of the features used to separate the two categories, but a planned detailed investigation will be able to reveal more details in future work. This will be based, for example, on layer-wise relevance propagation, a technique that can reveal the "grid cell relevance" of the temperature maps used as input for the output probabilities of a trained neural network (Bach et al., 2015; Toms et al., 2020; Labe and Barnes, 2022). For a comprehensive overview of such an approach applied to a related problem, including a review of the current literature, see, for example, Labe and Barnes (2022). These techniques investigating the classifiers themselves will be combined with background knowledge about climate models to provide integrated and interpretable insights into the origins of the classification skill.

In addition to the CMIP6 models, we also tested the classifiers on 1 year of prototype data from a global, storm-resolving simulation run with ICON-Sapphire at a resolution of 5 km. The logistic regression classifiers were unable to clearly assign the samples from ICON-Sapphire into either category, indicating that this high-resolution case has different and potentially reduced daily biases or more realistic covariance structure compared to the other, lower-resolution CMIP6 models. The CNN, in turn, managed to correctly predict that the samples from ICON-Sapphire with only the global mean removed belong to the model category for slightly more than 75% of samples, which is still a much lower value than for all other models. In the case where the seasonal cycle (estimated from only 1 year of data) was removed as well, the CNN misclassified the vast majority of samples as clearly belonging to the observation category.

This case raises a number of interesting questions that we will follow up upon once more data become available, for example: does the emerging higher similarity between ICON-Sapphire and observations truly reflect improved modeled characteristics of the daily temperature field covariance structure, or could it be due to some compensating error phenomena? If the former, is the higher similarity to observations due to reduced biases overall, or due to improved daily covariance structure of temperature fields? Which resolution or other improvements would it take to truly pass the "climate model Turing test" (Palmer, 2016) of inseparability of output fields of observations and climate models? The result for ICON-Sapphire is potentially highly encouraging, but we warrant that a cautious interpretation is needed, as only a short ICON-Sapphire simulation period was available at the writing of this study, and thus some effects of compensating errors cannot be definitely excluded.

https://doi.org/10.1017/eds.2023.23 Published online by Cambridge University Press

In the second part of the study, we investigated the ability of the CNN to identify each of the 43 models and the four observational datasets included in the study by their name. Again, we used daily temperature fields with the mean seasonal cycle and the daily global mean removed. We found an overall accuracy of 83%, which is about 40 times better than the baseline of a random choice. These results show that the CNN is clearly able to pick up relations between features that reliably separate models, including very similar variants from the same model families. At the same time, most of the misclassifications occur within model families or can be traced back to more distant "relations" such as common "ancestors." There is evidence that this behavior even holds for two of the reanalyses, ERA5 and MERRA2, which are misclassified as models from the EC-Earth and GFDL families, respectively, for more than 10% of cases and have documented common ancestry. Although more in-depth research is needed to confirm these results, they provide an interesting finding about the imprint of shared code even on short timescales and between development streams that have diverged many years ago.

Finally, we showed that the CNN is able to correctly identify about 70% of test samples even when they are drawn from the period 2091–2100 under the high-emission scenario SSP5-8.5, which is separated from the training period by about 100 years and features about 4 K warmer climate in the global average. Although this mean warming itself is not included in the samples, weather patterns are expected to change considerably due to climate change (Sippel et al., 2020). This means that the features used to identify climate models are—to a certain extent—state-invariant and thus remain robust even under a warming exceeding several degrees centigrade.

Future applications could build on the approaches illustrated here in several ways:

- They could add to the model evaluation toolbox (Eyring et al., 2019) and could target, for example, the classification of individual model components (e.g., atmospheric or ocean component), model generations (e.g., CMIP6 versus the roughly 10 years older CMIP5), perturbed parameter ensembles, or strains of model development in general. Following recent work from Labe and Barnes (2022), misclassifications could also be used systematically to draw conclusions about model performance.
- The classification approaches could be used to pinpoint model–model or model–observation differences and similarities. This could be done by analyzing the spatial scales of separability, that is, whether models on regional domains are less separable than globally, and/or whether this may depend on specific regions. Additionally, illustrating and understanding the patterns of separability, by using explainable neural network techniques (e.g., Toms et al., 2020), revealing how the neural network has learnt to distinguish models and observations, could provide additional insights.
- Related approaches could also be applied to infer model performance directly (e.g., predicting an estimate of climatological model error from short timescales) and possibly directly incorporating physical or dynamical processes (such as changes between two consecutive days) (Kashinath et al., 2021). This may be particularly relevant for high-resolution simulations such as the storm-resolving simulations currently in development, for which only short periods of simulation are available. In addition, recent efforts to merge climate predictions initialized from an observed state seamlessly with climate projections (Befort et al., 2022) may be interesting to analyze with our method, as it allows to diagnose from short timescales at which point the climatological model biases start to emerge, which may provide an opportunity for model development (Palmer, 2016).
- Here, we have focused on spatial patterns of mean temperature but other variables, such as precipitation, or other dimensions, such as the temporal distribution (potentially in a regional domain as discussed above) could also be considered to investigate their behavior.
- Our approach tests to which degree transferring patterns and relationships between models and observations is justified (Meinshausen, 2018). This assumption of distributional robustness is frequently made in the literature when classifiers are trained on simulated data and then applied to observed data (Gibson et al., 2021; Gordon et al., 2021; Kadow et al., 2020; Labe and Barnes, 2021; Sippel et al., 2021). From this perspective, our results can be seen as adversarial validation,

https://doi.org/10.1017/eds.2023.23 Published online by Cambridge University Press

which can be used to check whether the generalization from training to test sets is justified on different timescales (Shen et al., 2021).

- Lastly, and most speculatively, recent advances in machine learning towards image-to-image translation, using, for instance, techniques such as generative adversarial networks (Stengel et al., 2020), could provide an avenue to iteratively bias-correct model output in relation to observations, until a hypothetical "bias-corrected" spatial pattern would indistinguishable from observations. The idea in such generative adversarial approaches is precisely that a classifier cannot tell the difference between a simulated, "bias-corrected" output field, and an observed one.

Acknowledgments. The authors thank Reto Knutti, Erich M. Fischer and Christoph Schär (all at ETH Zurich), Aiko Voigt (University of Vienna), Bjorn Stevens (Max Planck Institute for Meteorology, Hamburg), and Michael Notter (EPFL, Lausanne) for valuable discussions and feedback on various aspects of this paper. We thank two anonymous reviewers for their helpful comments. We acknowledge the World Climate Research Programme, which, through its Working Group on Coupled Modeling, coordinated and promoted CMIP6. We thank the climate modeling groups for producing and making their model output available, the Earth System Grid Federation (ESGF) for archiving the data and providing access, and the multiple funding agencies that support CMIP6 and ESGF. This study was generated using Copernicus Climate Change Service information from ERA5. Support for the 20CR dataset used in this work is provided by the US Department of Energy, Office of Science Biological and Environmental Research (BER), by the National Oceanic and Atmospheric Administration Climate Program Office, and by the NOAA Physical Sciences Laboratory. The authors thank the nextGEMS project for providing the storm-resolving ICON-Sapphire runs, the NASA GMAO for providing MERRA2, NOAA for providing DOISST, and Urs Beyerle for downloading the CMIP6 data used in this work.

Author contribution. This work was conceptualized and written by L.B. with contributions from S.S. Data acquisition, analysis, and visualization by L.B. Both authors approved the final submitted draft.

Competing interest. The authors declare none.

Data availability statement. All data used in this study are freely available for research applications. Lists of all used CMIP6 models and observations are included in the Supplementary Tables S1–S3.

Ethics statement. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

Funding statement. This research was supported by the H2020 European Research Council project "European Climate Prediction System" (EUCP; grant no. 776613). S.S. acknowledges funding received from the Swiss National Science Foundation within the project "Combining theory with Big Data? The case of uncertainty in prediction of trends in extreme weather and impacts" (grant no. 167215), the Swiss Data Science Centre within the project "Data Science-informed attribution of changes in the Hydrological cycle" (DASH; C17–01) and within the European Union H2020 project "Artificial intelligence for detection and attribution" (XAIDA; grant no. 101003469).

Provenance statement. This article is part of the Climate Informatics 2023 proceedings and was accepted in *Environmental Data Science* on the basis of the Climate Informatics peer review process.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/eds.2023.23.

References

Andrews T, **Andrews MB**, **Bodas-Salcedo A**, **Jones GS**, **Kuhlbrodt T**, **Manners J**, **Menary MB**, **Ridley J**, **Ringer MA**, **Sellar AA**, **Senior CA and Tang Y** (2019) Forcings, feedbacks, and climate sensitivity in HadGEM3-GC3.1 and UKESM1. *Journal of Advances in Modeling Earth Systems 11*(12), 4377–4394. https://doi.org/10.1029/2019MS001866

Annan JD and Hargreaves JC (2017) On the meaning of independence in climate science. *Earth System Dynamics 8*(1), 211–224. https://doi.org/10.5194/esd-8-211-2017

Bach S, **Binder A**, **Montavon G**, **Klauschen F**, **Müller KR and Samek W** (2015) On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One 10*(7), 1–46. https://doi.org/10.1371/journal.pone.0130140

Barnes EA, **Hurrell JW**, **Ebert-Uphoff I**, **Anderson C and Anderson D** (2019) Viewing forced climate patterns through an AI lens. *Geophysical Research Letters 46*(22), 13389–13398. https://doi.org/10.1029/2019GL084944

Bauer P, **Stevens B and Hazeleger W** (2021) A digital twin of earth for the green transition. *Nature Climate Change 11*(2), 80–83. https://doi.org/10.1038/s41558-021-00986-y

Befort DJ, **Brunner L**, **Borchert LF**, **O'Reilly CH**, **Mignot J**, **Ballinger AP**, **Hegerl GC**, **Murphy JM and Weisheimer A** (2022) Combination of decadal predictions and climate projections in time: Challenges and potential solutions. *Geophysical Research Letters 49*(15), 1–18. https://doi.org/10.1029/2022GL098568

Bi D, **Dix M**, **Marsland S**, **O'farrell S**, **Sullivan A**, **Bodman R**, **Law R**, **Harman I**, **Srbinovsky J**, **Rashid HA**, **Dobrohotoff P**, **Mackallah C**, **Yan H**, **Hirst A**, **Savita A**, **Dias FB**, **Woodhouse M**, **Fiedler R and Heerdegen A** (2020) Configuration and spin-up of ACCESS-CM2, the new generation Australian Community climate and earth system simulator coupled model. *Journal of Southern Hemisphere Earth Systems Science 70*(1), 225–251. https://doi.org/10.1071/ES19040

Bishop CH and Abramowitz G (2013) Climate model dependence and the replicate earth paradigm. *Climate Dynamics 41*(3–4), 885–900. https://doi.org/10.1007/s00382-012-1610-y

Bock L, **Lauer A**, **Schlund M**, **Barreiro M**, **Bellouin N**, **Jones C**, **Meehl GA**, **Predoi V**, **Roberts MJ and Eyring V** (2020) Quantifying progress across different CMIP phases with the ESMValTool. *Journal of Geophysical Research – Atmospheres 125* (21), 1–28. https://doi.org/10.1029/2019JD032321

Boé J (2018) Interdependency in multimodel climate projections: Component replication and result similarity. *Geophysical Research Letters 45*(6), 2771–2779. https://doi.org/10.1002/2017GL076829

Brient F, **Roehrig R and Voldoire A** (2019) Evaluating marine stratocumulus clouds in the CNRM-CM6-1 model using short-term Hindcasts. *Journal of Advances in Modeling Earth Systems 11*(1), 127–148. https://doi.org/10.1029/2018MS001461

Brunner L, **Pendergrass AG**, **Lehner F**, **Merrifield AL**, **Lorenz R and Knutti R** (2020) Reduced global warming from CMIP6 projections when weighting models by performance and independence. *Earth System Dynamics 11*(4), 995–1012. https://doi.org/10.5194/esd-11-995-2020

Brunner L and Sippel S (2023) Data and code for "identifying climate models based on their daily output using machine learning". *Zenodo*. https://doi.org/10.5281/zenodo.7998436

Chemke R, **Zanna L and Polvani LM** (2020) Identifying a human signal in the North Atlantic warming hole. *Nature Communications 11*(1), 1–7. https://doi.org/10.1038/s41467-020-15285-x

Chen G, **Wang W-C**, **Bao Q and Li J** (2022) Evaluation of simulated cloud diurnal variation in CMIP6 climate models. *Journal of Geophysical Research – Atmospheres 127*(6), 1–14. https://doi.org/10.1029/2021jd036422

Cherchi A, **Fogli PG**, **Lovato T**, **Peano D**, **Iovino D**, **Gualdi S**, **Masina S**, **Scoccimarro E**, **Materia S**, **Bellucci A and Navarra A** (2019) Global mean climate and main patterns of variability in the CMCC-CM2 coupled model. *Journal of Advances in Modeling Earth Systems 11*(1), 185–209. https://doi.org/10.1029/2018MS001369

Davy R and Outten S (2020) The Arctic surface climate in CMIP6: Status and developments since CMIP5. *Journal of Climate 33* (18), 8047–8068. https://doi.org/10.1175/JCLI-D-19-0990.1

de Vries IE, **Sippel S**, **Pendergrass AG and Knutti R** (2023) Robust global detection of forced changes in mean and extreme precipitation despite observational disagreement on the magnitude of change. *Earth System Dynamics 14*(1), 81–100. https://doi.org/10.5194/esd-14-81-2023

Döscher R, **Acosta M**, **Alessandri A**, **Anthoni P**, **Arsouze T**, **Bergman T**, **Bernardello R**, **Boussetta S**, **Caron L-P**, **Carver G**, **Castrillo M**, **Catalano F**, **Cvijanovic I**, **Davini P**, **Dekker E**, **Doblas-Reyes FJ**, **Docquier D**, **Echevarria P**, **Fladrich U**, **Fuentes-Franco R**, **Gröger M**, **Hardenberg JV**, **Hieronymus J**, **Karami MP**, **Keskinen J-P**, **Koenigk T**, **Makkonen R**, **Massonnet F**, **Ménégoz M**, **Miller PA**, **Moreno-Chamarro E**, **Nieradzik L**, **Noije Tv**, **Nolan P**, **O'Donnell D**, **Ollinaho P**, **Oord Gvd**, **Ortega P**, **Prims OT**, **Ramos A**, **Reerink T**, **Rousset C**, **Ruprich-Robert Y**, **Sager PL**, **Schmith T**, **Schrödner R**, **Serva F**, **Sicardi V**, **Madsen MS**, **Smith B**, **Tian T**, **Tourigny E**, **Uotila P**, **Vancoppenolle M**, **Wang S**, **Wårlind D**, **Willén U**, **Wyser K**, **Yang S**, **Yepes-Arbós X and Zhang Q** (2022) The EC-Earth3 earth system model for the coupled model Intercomparison project 6. *Geoscientific Model Development 15*(7), 2973–3020. https://doi.org/10.5194/gmd-15-2973-2022

Dunne JP, **Horowitz LW**, **Adcroft AJ**, **Ginoux P**, **Held IM**, **John JG**, **Krasting JP**, **Malyshev S**, **Naik V**, **Paulot F**, **Shevliakova E**, **Stock CA**, **Zadeh N**, **Balaji V**, **Blanton C**, **Dunne KA**, **Dupuis C**, **Durachta J**, **Dussin R**, **Gauthier PPG**, **Griffies SM**, **Guo H**, **Hallberg RW**, **Harrison M**, **He J**, **Hurlin W**, **McHugh C**, **Menzel R**, **Milly PCD**, **Nikonov S**, **Paynter DJ**, **Ploshay J**, **Radhakrishnan A**, **Rand K**, **Reichl BG**, **Robinson T**, **Schwarzkopf DM**, **Sentman LT**, **Underwood S**, **Vahlenkamp H**, **Winton M**, **Wittenberg AT**, **Wyman B**, **Zeng Y and Zhao M** (2020) The GFDL earth system model version 4.1 (GFDL-ESM 4.1): Overall coupled model description and simulation characteristics. *Journal of Advances in Modeling Earth Systems 12*(11), 1–56. https://doi.org/10.1029/2019MS002015

Eyring V, **Bony S**, **Meehl GA**, **Senior CA**, **Stevens B**, **Stouffer RJ and Taylor KE** (2016) Overview of the coupled model Intercomparison project phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development 9*(5), 1937–1958. https://doi.org10.5194/gmd-9-1937-2016

Eyring V, **Cox PM**, **Flato GM**, **Gleckler PJ**, **Abramowitz G**, **Caldwell P**, **Collins WD**, **Gier BK**, **Hall AD**, **Hoffman FM**, **Hurtt GC**, **Jahn A**, **Jones CD**, **Klein SA**, **Krasting JP**, **Kwiatkowski L**, **Lorenz R**, **Maloney E**, **Meehl GA**, **Pendergrass AG**, **Pincus R**, **Ruane AC**, **Russell JL**, **Sanderson BM**, **Santer BD**, **Sherwood SC**, **Simpson IR**, **Stouffer RJ and Williamson MS** (2019) Taking climate model evaluation to the next level. *Nature Climate Change 9*(2), 102–110. https://doi.org/10.1038/s41558-018-0355-y

Gelaro R, **McCarty W**, **Suárez MJ**, **Todling R**, **Molod A**, **Takacs L**, **Randles CA**, **Darmenov A**, **Bosilovich MG**, **Reichle R**, **Wargan K**, **Coy L**, **Cullather R**, **Draper C**, **Akella S**, **Buchard V**, **Conaty A**, **da Silva AM**, **Gu W**, **Kim GK**, **Koster R**, **Lucchesi R**, **Merkova D**, **Nielsen JE**, **Partyka G**, **Pawson S**, **Putman W**, **Rienecker M**, **Schubert SD**, **Sienkiewicz M and Zhao B** (2017) The modern-era retrospective analysis for research and applications, version 2 (MERRA-2). *Journal of Climate 30*(14), 5419–5454. https://doi.org/10.1175/JCLI-D-16-0758.1

Gibson PB, **Chapman WE**, **Altinok A**, **Monache LD**, **DeFlorio MJ and Waliser DE** (2021) Training machine learning models on climate model output yields skillful interpretable seasonal precipitation forecasts. *Communications Earth & Environment 2*(1), 159. https://doi.org/10.1038/s43247-021-00225-4

https://doi.org/10.1017/eds.2023.23 Published online by Cambridge University Press

Giorgi F and Coppola E (2010) Does the model regional bias affect the projected regional climate change? An analysis of global model projections: A letter. *Climatic Change 100*(3), 787–795. https://doi.org/10.1007/s10584-010-9864-z

GMAO (2015) MERRA-2 tavg1_2d_slv_Nx: 2d,1-Hourly,Time-Averaged,Single-Level,Assimilation,Single-Level Diagnostics V5.12.4.

Gordon EM, Barnes EA and Hurrell JW (2021) Oceanic harbingers of Pacific decadal oscillation predictability in CESM2 detected by neural networks. *Geophysical Research Letters 48*(21), 0094–8276. https://doi.org/10.1029/2021gl095392

Hastie T, Tibshirani R and Friedman J (2009) *The Elements of Statistical Learning*, 2nd Edn. New York, NY: Springer. https://doi.org/10.1080/01443610062940

Hersbach H, Bell B, Berrisford P, Hirahara S, Horányi A, Muñoz-Sabater J, Nicolas J, Peubey C, Radu R, Schepers D, Simmons A, Soci C, Abdalla S, Abellan X, Balsamo G, Bechtold P, Biavati G, Bidlot J, Bonavita M, Chiara GD, Dahlgren P, Dee D, Diamantakis M, Dragani R, Flemming J, Forbes R, Fuentes M, Geer A, Haimberger L, Healy S, Hogan RJ, Hólm E, Janisková M, Keeley S, Laloyaux P, Lopez P, Lupu C, Radnoti G, Rosnay Pd, Rozum I, Vamborg F, Villaume S and Thépaut JN (2020) The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society 146*(730), 1999–2049. https://doi.org/10.1002/qj.3803

Hirota N, Takayabu YN, Watanabe M and Kimoto M (2011) Precipitation reproducibility over tropical oceans and its relationship to the double ITCZ problem in CMIP3 and MIROC5 climate models. *Journal of Climate 24*(18), 4859–4873. https://doi.org/10.1175/2011JCLI4156.1

Hohenegger C, Korn P, Linardakis L, Redler R, Schnur R, Adamidis P, Bao J, Bastin S, Behravesh M, Bergemann M, Biercamp J, Bockelmann H, Brokopf R, Brüggemann N, Casaroli L, Chegini F, Datseris G, Esch M, George G, Giorgetta M, Gutjahr O, Haak H, Hanke M, Ilyina T, Jahns T, Jungclaus J, Kern M, Klocke D, Kluft L, Kölling T, Kornblueh L, Kosukhin S, Kroll C, Lee J, Mauritsen T, Mehlmann C, Mieslinger T, Naumann AK, Paccini L, Peinado A, Praturi DS, Putrasahan D, Rast S, Riddick T, Roeber N, Schmidt H, Schulzweida U, Schütte F, Segura H, Shevchenko R, Singh V, Specht M, Stephan CC, Vogel R, Wengel C, Winkler M, Ziemen F, Marotzke J and Stevens B (2023) ICON-Sapphire: Simulating the components of the Earth System and their interactions at kilometer and subkilometer scales. *Geoscientific Model Development 16*, 779–811. https://doi.org/10.5194/gmd-2022-171

Hsi Yen Ma SX, Klein SA, Williams KD, Boyle JS, Bony S, Douville H, Fermepin S, Medeiros B, Tyteca S, Watanabe M and Williamson D (2014) On the correspondence between mean forecast errors and climate errors in CMIP5 models. *Journal of Climate 27*(4), 1781–1798. https://doi.org/10.1175/JCLI-D-13-00474.1

Hsieh WW (2022) Evolution of machine learning in environmental science—A perspective. *Environmental Data Science 1*, 1–8. https://doi.org/10.1017/eds.2022.2

Huang B, Liu C, Banzon V, Freeman E, Graham G, Hankins B, Smith T and Zhang HM (2021) Improvements of the daily optimum Interpolation Sea surface temperature (DOISST) version 2.1. *Journal of Climate 34*(8), 2923–2939. https://doi.org/10.1175/JCLI-D-20-0166.1

IPCC (2021) Summary for policymakers. In Masson-Delmotte, V., P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou (eds.) *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, UK and New York, NY, USA: Cambridge University Press.

Kadow C, Hall DM and Ulbrich U (2020) Artificial intelligence reconstructs missing climate information. *Nature Geoscience 13*(6), 408–413. https://doi.org/10.1038/s41561-020-0582-5

Kashinath K, Mustafa M, Albert A, Wu JL, Jiang C, Esmaeilzadeh S, Azizzadenesheli K, Wang R, Chattopadhyay A, Singh A, Manepalli A, Chirila D, Yu R, Walters R, White B, Xiao H, Tchelepi HA, Marcus P, Anandkumar A, Hassanzadeh P, and Prabhat (2021) Physics-informed machine learning: Case studies for weather and climate modelling. *Philosophical Transactions of the Royal Society A - Mathematical Physical and Engineering Sciences 379*(2194). https://doi.org/10.1098/rsta.2020.0093

Keil P, Mauritsen T, Jungclaus J, Hedemann C, Olonscheck D and Ghosh R (2020) Multiple drivers of the North Atlantic warming hole. *Nature Climate Change 10*(7), 667–671. https://doi.org/10.1038/s41558-020-0819-8

Knutti R (2010) The end of model democracy? *Climatic Change 102*(3–4), 395–404. https://doi.org/10.1007/s10584-010-9800-2

Knutti R, Masson D and Gettelman A (2013) Climate model genealogy: Generation CMIP5 and how we got there. *Geophysical Research Letters 40*(6), 1194–1199. https://doi.org/10.1002/grl.50256

Labe ZM and Barnes EA (2021) Detecting climate signals using explainable AI with single-forcing large ensembles. *Journal of Advances in Modeling Earth Systems 13*(6), 1–18. https://doi.org/10.1029/2021MS002464

Labe ZM and Barnes EA (2022) Comparison of climate model large ensembles with observations in the Arctic using simple neural networks. *Earth and Space Science 9*(7), 1–18. https://doi.org/10.1029/2022EA002348

Lee J, Kim JJ, Sun MA, Kim BH, Moon H, Sung HM, Kim JJ and Byun YH (2020) Evaluation of the Korea Meteorological Administration advanced Community earth-system model (K-ACE). *Asia-Pacific Journal of Atmospheric Sciences 56*(3), 381–395. https://doi.org/10.1007/s13143-019-00144-7

Li G and Xie SP (2014) Tropical biases in CMIP5 multimodel ensemble: The excessive equatorial pacific cold tongue and double ITCZ problems. *Journal of Climate 27*(4), 1765–1780. https://doi.org/10.1175/JCLI-D-13-00337.1

Lin SJ (2004) A "vertically Lagrangian" finite-volume dynamical core for global models. *Monthly Weather Review 132*(10), 2293–2307. https://doi.org/10.1175/1520-0493(2004)132<2293:AVLFDC>2.0.CO;2

https://doi.org/10.1017/eds.2023.23 Published online by Cambridge University Press

Masson D and Knutti R (2011) Climate model genealogy. *Geophysical Research Letters 38*(8), 1–4. https://doi.org/10.1029/2011GL046864

Mauritsen T, **Bader J**, **Becker T**, **Behrens J**, **Bittner M**, **Brokopf R**, **Brovkin V**, **Claussen M**, **Crueger T**, **Esch M**, **Fast I**, **Fiedler S**, **Fläschner D**, **Gayler V**, **Giorgetta M**, **Goll DS**, **Haak H**, **Hagemann S**, **Hedemann C**, **Hohenegger C**, **Ilyina T**, **Jahns T**, **Jimenéz-de-la Cuesta D**, **Jungclaus J**, **Kleinen T**, **Kloster S**, **Kracher D**, **Kinne S**, **Kleberg D**, **Lasslop G**, **Kornblueh L**, **Marotzke J**, **Matei D**, **Meraner K**, **Mikolajewicz U**, **Modali K**, **Möbis B**, **Müller WA**, **Nabel JEMS**, **Nam CCW**, **Notz D**, **Nyawira SS**, **Paulsen H**, **Peters K**, **Pincus R**, **Pohlmann H**, **Pongratz J**, **Popp M**, **Raddatz TJ**, **Rast S**, **Redler R**, **Reick CH**, **Rohrschneider T**, **Schemann V**, **Schmidt H**, **Schnur R**, **Schulzweida U**, **Six KD**, **Stein L**, **Stemmler I**, **Stevens B**, **von Storch JS**, **Tian F**, **Voigt A**, **Vrese P**, **Wieners KH**, **Wilkenskjeld S**, **Winkler A and Roeckner E** (2019) Developments in the MPI-M earth system model version 1.2 (MPI-ESM1.2) and its response to increasing CO2. *Journal of Advances in Modeling Earth Systems 11*(4), 998–1038. https://doi.org/10.1029/2018MS001400

Meehl GA, **Richter JH**, **Teng H**, **Capotondi A**, **Cobb K**, **Doblas-Reyes F**, **Donat MG**, **England MH**, **Fyfe JC**, **Han W**, **Kim H**, **Kirtman BP**, **Kushnir Y**, **Lovenduski NS**, **Mann ME**, **Merryfield WJ**, **Nieves V**, **Pegion K**, **Rosenbloom N**, **Sanchez SC**, **Scaife AA**, **Smith D**, **Subramanian AC**, **Sun L**, **Thompson D**, **Ummenhofer CC and Xie SP** (2021) Initialized earth system prediction from subseasonal to decadal timescales. *Nature Reviews Earth & Environment 2*(5), 340–357. https://doi.org/10.1038/s43017-021-00155-x

Meinshausen N (2018) Causality from a distributional robustness point of view. In *IEEE Data Science Workshop (DSW)*. Lausanne, Switzerland: IEEE, pp. 6–10. https://doi.org/10.1109/DSW.2018.8439889

Meinshausen M, **Nicholls ZRJ**, **Lewis J**, **Gidden MJ**, **Vogel E**, **Freund M**, **Beyerle U**, **Gessner C**, **Nauels A**, **Bauer N**, **Canadell JG**, **Daniel JS**, **John A**, **Krummel PB**, **Luderer G**, **Meinshausen N**, **Montzka SA**, **Rayner PJ**, **Reimann S**, **Smith SJ**, **van den Berg M**, **Velders GJM**, **Vollmer MK and Wang RHJ** (2020) The shared socio-economic pathway (SSP) greenhouse gas concentrations and their extensions to 2500. *Geoscientific Model Development 13*(8), 3571–3605. https://doi.org/10.5194/gmd-13-3571-2020

Merrifield AL, **Brunner L**, **Lorenz R**, **Humphrey V and Knutti R** (2023) Climate model Selection by Independence, Performance, and Spread (ClimSIPS) for regional applications. (February):1–49. https://doi.org/10.5194/egusphere-2022-1520

Merrifield AL, **Brunner L**, **Lorenz R**, **Medhaug I and Knutti R** (2020) An investigation of weighting schemes suitable for incorporating large ensembles into multi-model ensembles. *Earth System Dynamics 11*(3), 807–834. https://doi.org/10.5194/esd-11-807-2020

Molod A, **Takacs L**, **Suarez M and Bacmeister J** (2015) Development of the GEOS-5 atmospheric general circulation model: Evolution from MERRA to MERRA2. *Geoscientific Model Development 8*(5), 1339–1356. https://doi.org/10.5194/gmd-8-1339-2015

Notz D and Community S (2020) Arctic Sea ice in CMIP6. *Geophysical Research Letters 47*(10), 1–11. https://doi.org/10.1029/2019GL086749

Palmer TN (2016) A personal perspective on modelling the climate system. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 472*(2188), 20150772. https://doi.org/10.1098/rspa.2015.0772

Rackow T, **Wedi N**, **Mogensen K**, **Dueben P**, **Goessling HF**, **Hegewald J**, **Kühnlein C**, **Zampieri L and Jung T** (2021) DYAMOND-II simulations with IFS-FESOM2, EGU General Assembly 2021, online, 19–30 April 2021, EGU21-9672. https://doi.org/10.5194/egusphere-egu21-9672

Rienecker MM, **Suarez MJ**, **Todling R**, **Bacmeister J**, **Takacs L**, **Liu H-C**, **Gu W**, **Sienkiewicz M**, **Koster RD**, **Gelaro R**, **Stajner I and Nielsen JE** (2008) The GEOS-5 Data Assimilation System—Documentation of versions 5.0.1 and 5.1.0, and 5.2.0. Technical Report Series on Global Modeling and Data Assimilation, Volume 27. NASA/TM-2008-104606, 27(December), 92 pp.

Shen Z, **Liu J**, **He Y**, **Zhang X**, **Xu R**, **Yu H and Cui P** (2021) Towards out-of-distribution generalization: A survey. *Journal of Latex Class Files 14*(8), 1–22. Available at http://arxiv.org/abs/2108.13624.

Sippel S, **Meinshausen N**, **Fischer EM**, **Székely E and Knutti R** (2020) Climate change now detectable from any single day of weather at global scale. *Nature Climate Change 10*(1), 35–41. https://doi.org/10.1038/s41558-019-0666-7

Sippel S, **Meinshausen N**, **Székely E**, **Fischer E**, **Pendergrass AG**, **Lehner F and Knutti R** (2021) Robust detection of forced warming in the presence of potentially large climate variability. *Science Advances 7*(43), 1–18. https://doi.org/10.1126/sciadv.abh4429

Slivinski LC, **Compo GP**, **Sardeshmukh PD**, **Whitaker JS**, **McColl C**, **Allan RJ**, **Brohan P**, **Yin X**, **Smith CA**, **Spencer LJ**, **Vose RS**, **Rohrer M**, **Conroy RP**, **Schuster DC**, **Kennedy JJ**, **Ashcroft L**, **Brönnimann S**, **Brunet M**, **Camuffo D**, **Cornes R**, **Cram TA**, **Domínguez-Castro F**, **Freeman JE**, **Gergis J**, **Hawkins E**, **Jones PD**, **Kubota H**, **Lee TC**, **Lorrey AM**, **Luterbacher J**, **Mock CJ**, **Przybylak RK**, **Pudmenzky C**, **Slonosky VC**, **Tinz B**, **Trewin B**, **Wang XL**, **Wilkinson C**, **Wood K and Wyszynski P** (2021) An evaluation of the performance of the twentieth century reanalysis version 3. *Journal of Climate 34*(4), 1417–1438. https://doi.org/10.1175/JCLI-D-20-0505.1

Stengel K, **Glaws A**, **Hettinger D and King RN** (2020) Adversarial super-resolution of climatological wind and solar data. *Proceedings of the National Academy of Sciences of the United States of America 117*(29), 16805–16815. https://doi.org/10.1073/pnas.1918964117

Tebaldi C and Knutti R (2007) The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions of the Royal Society A - Mathematical Physical and Engineering Sciences 365*(1857), 2053–2075. https://doi.org/10.1098/rsta.2007.2076

https://doi.org/10.1017/eds.2023.23 Published online by Cambridge University Press

Tian B and Dong X (2020) The double-ITCZ bias in CMIP3, CMIP5, and CMIP6 models based on annual mean precipitation. *Geophysical Research Letters 470*(8), 1–11. https://doi.org/10.1029/2020GL087232

Toms BA, **Barnes EA and Ebert-Uphoff I** (2020) Physically interpretable neural networks for the geosciences: Applications to earth system variability. *Journal of Advances in Modeling Earth Systems 120*(9), 1–20. https://doi.org/10.1029/2019MS002002

Williams KD, **Bodas-Salcedo A**, **Déqué M**, **Fermepin S**, **Medeiros B**, **Watanabe M**, **Jakob C**, **Klein SA**, **Senior CA and Williamson DL** (2013) The transpose-AMIP II experiment and its application to the understanding of southern ocean cloud biases in climate models. *Journal of Climate 260*(10), 3258–3274. https://doi.org/10.1175/JCLI-D-12-00429.1

Cite this article: Brunner L and Sippel S (2023). Identifying climate models based on their daily output using machine learning. *Environmental Data Science*, 2: e22. doi:10.1017/eds.2023.23