Improving the reliability of ML-corrected climate models with novelty detection

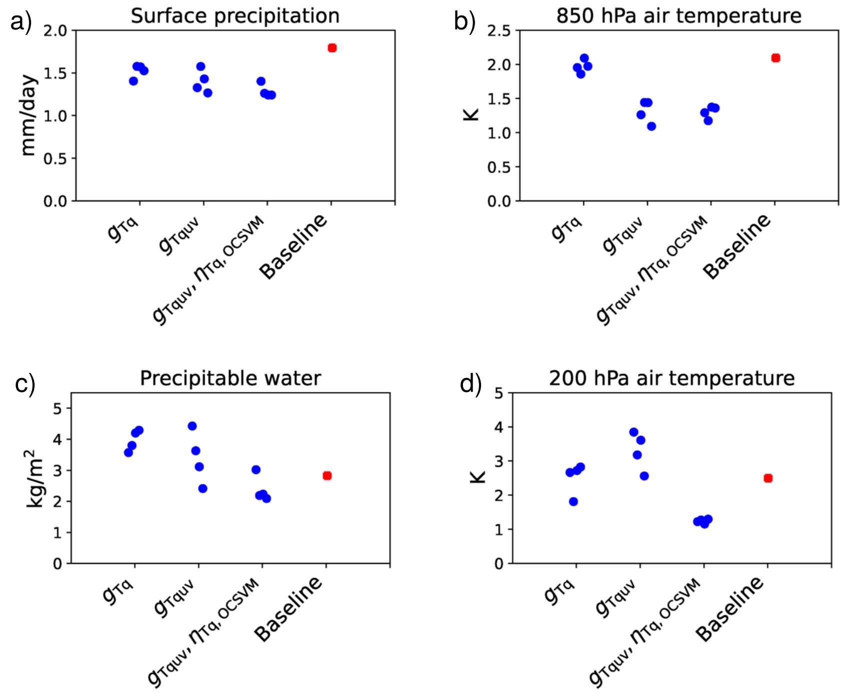
Clayton Hendrick Sanford[1], Anna Kwa[2], Oliver Watt-Meyer[2], Spencer Koncius Clark[3], Noah Domino Brenowitz[4], Jeremy McGibbon[2], and Christopher S. Bretherton[2]

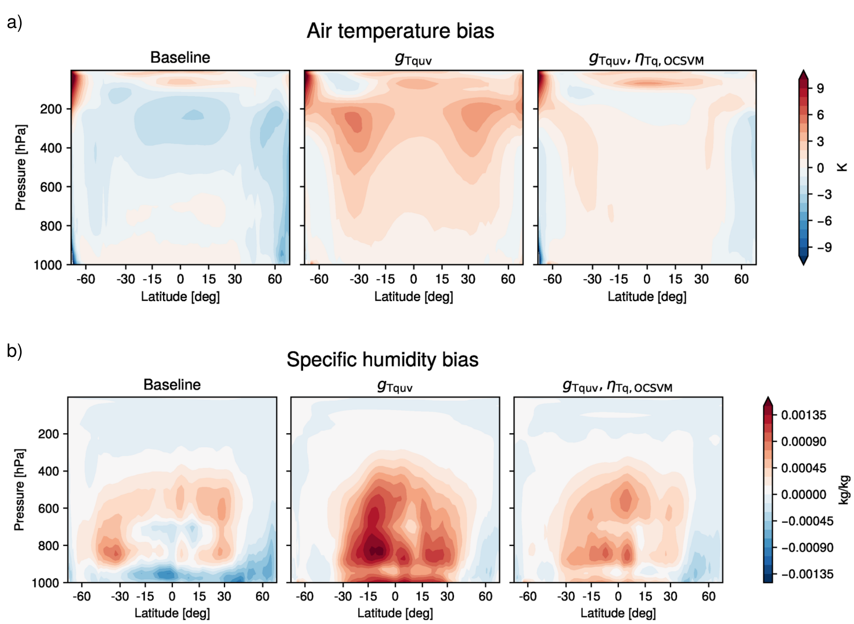
[1]Columbia University
[2]Allen Institute for Artificial Intelligence
[3]Allen Institute for Artificial Intelligence / NOAA-GFDL
[4]NVIDIA

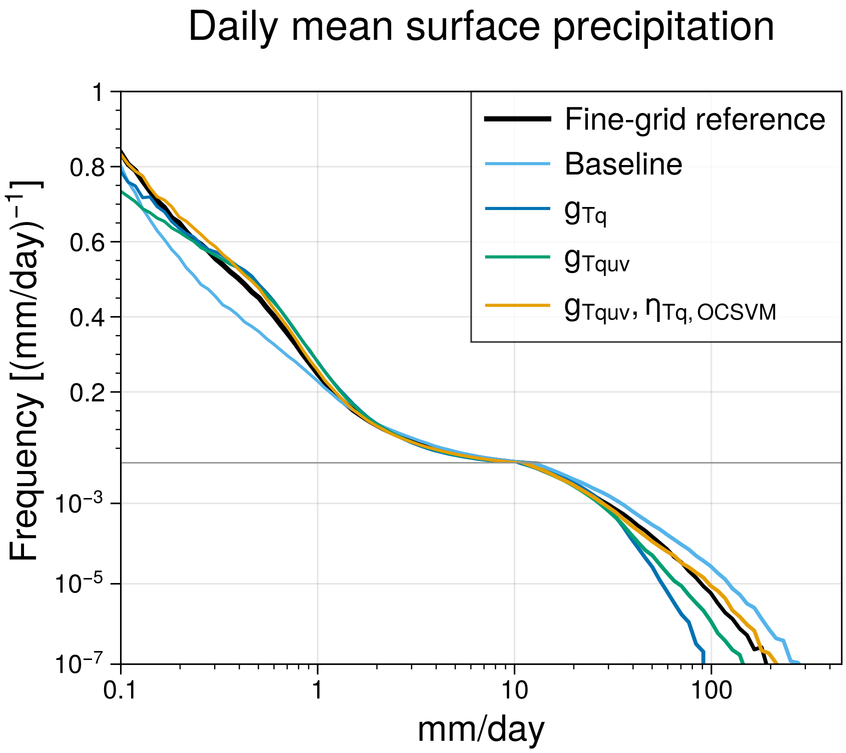
May 25, 2023

Abstract

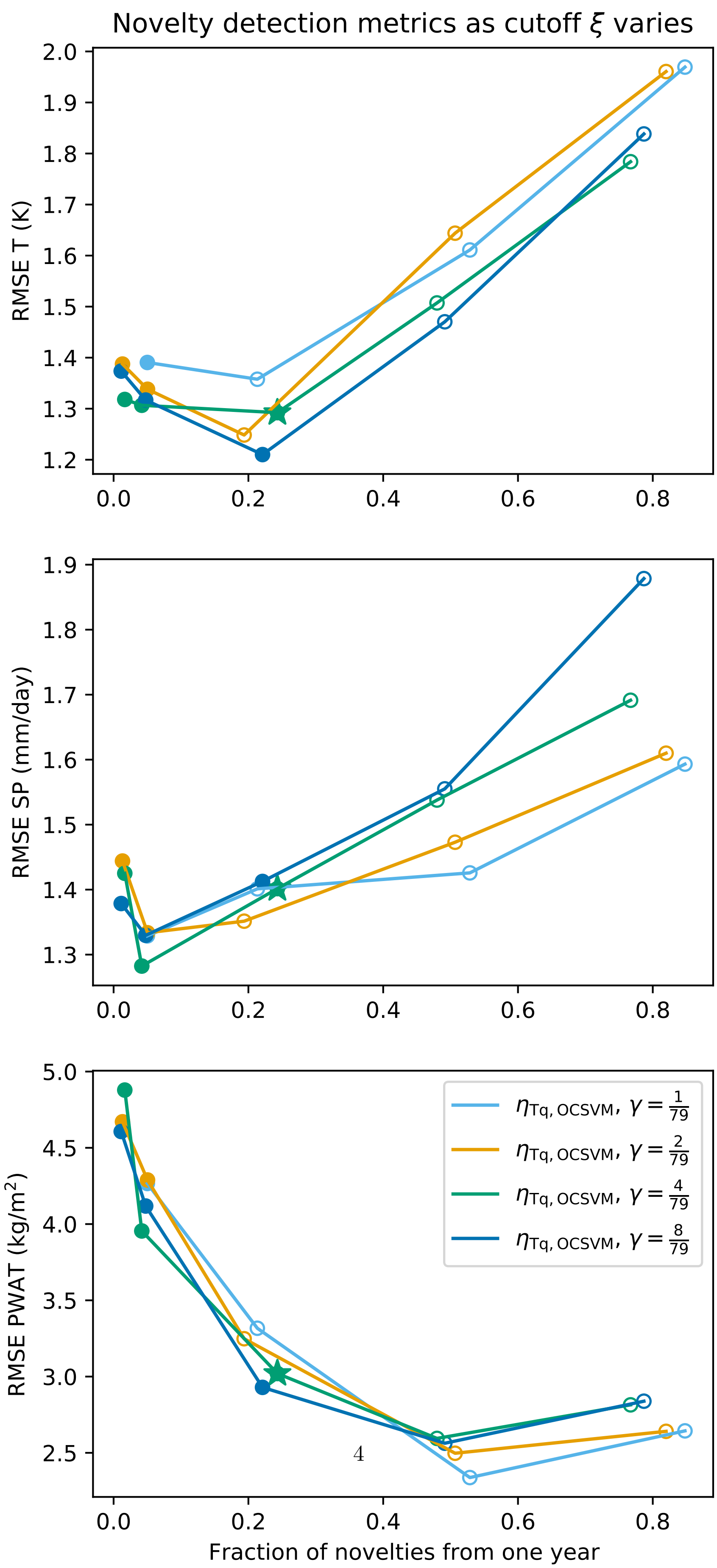
The use of machine learning (ML) for the online correction of coarse-resolution atmospheric models has proven effective in reducing biases in near-surface temperature and precipitation rate. However, this often introduces biases in the upper atmosphere and improvements are not always reliable across ML-corrective models trained with different random seeds. Furthermore, ML corrections can feed back on the baseline physics of the atmospheric model and produce profiles that are outside the distribution of samples used in training, leading to low confidence in the predicted corrections. This study introduces the use of a novelty detector to mask the predicted corrections when the atmospheric state is deemed out-of-sample. The novelty detector is trained on profiles of temperature and specific humidity in a semi-supervised fashion using samples from the coarsened reference fine-resolution simulation. Offline, the novelty detector determines more columns to be out-of-sample in simulations which are known, using simple metrics like mean bias, to drift further from the reference simulation. Without novelty detection, corrective ML leads to the development of undesirably large climate biases for some ML random seeds but not others. Novelty detection deems about 21% of columns to be novelties in year-long simulations. The spread in the root mean square error (RMSE) of time-mean spatial patterns of surface temperature and precipitation rate across a random seed ensemble is sharply reduced when using novelty detection. In particular, the random seed with the worst RMSE is improved by up to 60% (depending on the variable) while the best seed maintains its low RMSE.


a)
Surface precipitation
mm/day
b)
850 hPa air temperature
K
c)
Precipitable water
kg/m²
d)
200 hPa air temperature
K
Baseline

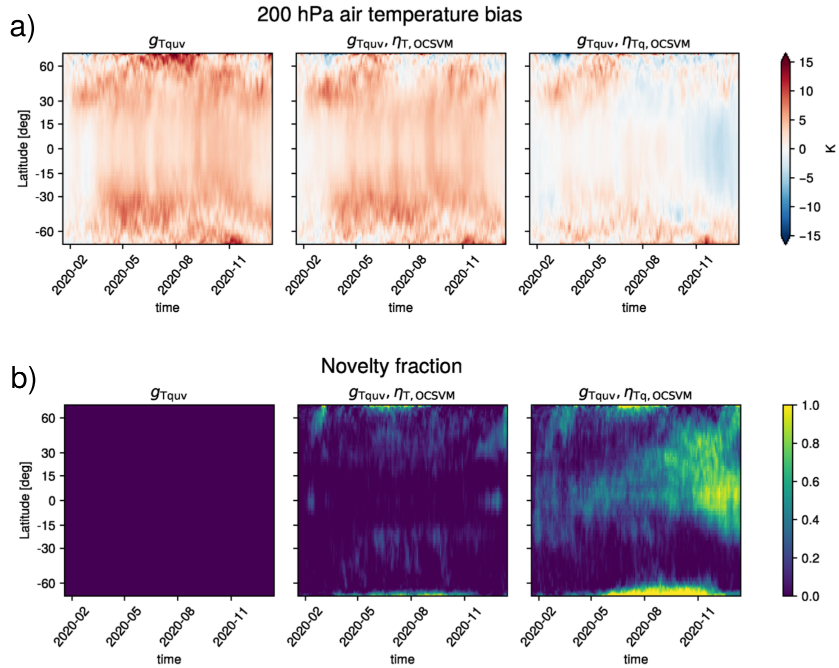

a)
Air temperature bias
Baseline
Pressure [hPa]
Latitude [deg]
K
b)
Specific humidity bias
Baseline
Pressure [hPa]
Latitude [deg]
kg/kg

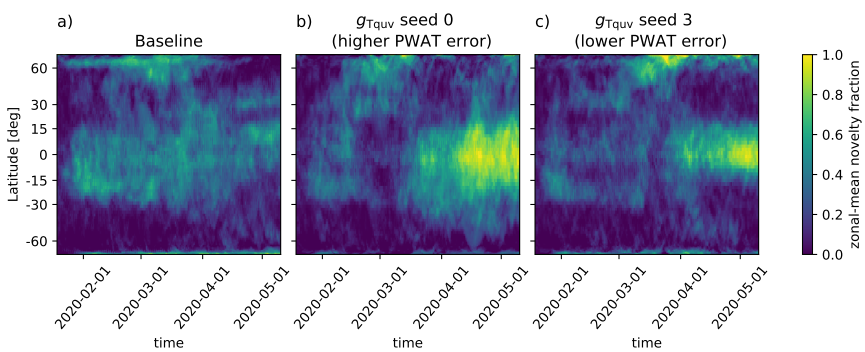


Daily mean surface precipitation
Frequency [(mm/day)$^{-1}$]
1
0.8
0.6
0.4
0.2
10^{-3}
10^{-5}
10^{-7}
0.1
1
10
100
mm/day
Fine-grid reference
Baseline
g_{Tq}
g_{Tquv}
$g_{Tquv}, \eta_{Tq,\mathrm{OCSVM}}$



Novelty detection metrics as cutoff ξ varies
RMSE T (K)
RMSE SP (mm/day)
RMSE PWAT (kg/m^2)
Fraction of novelties from one year
$\eta_{\mathrm{Tq,OCSVM}}$, $\gamma = \frac{1}{79}$
$\eta_{\mathrm{Tq,OCSVM}}$, $\gamma = \frac{2}{79}$
$\eta_{\mathrm{Tq,OCSVM}}$, $\gamma = \frac{4}{79}$
$\eta_{\mathrm{Tq,OCSVM}}$, $\gamma = \frac{8}{79}$


a)
200 hPa air temperature bias
g_{Tquv}
$g_{\mathrm{Tquv}}, \eta_{\mathrm{T,OCSVM}}$
$g_{\mathrm{Tquv}}, \eta_{\mathrm{Tq,OCSVM}}$
Latitude [deg]
time
K
b)
Novelty fraction
g_{Tquv}
$g_{\mathrm{Tquv}}, \eta_{\mathrm{T,OCSVM}}$
$g_{\mathrm{Tquv}}, \eta_{\mathrm{Tq,OCSVM}}$
Latitude [deg]
time


a)
Baseline
b)
g_{Tquv} seed 0
(higher PWAT error)
c)
g_{Tquv} seed 3
(lower PWAT error)
Latitude [deg]
time
zonal-mean novelty fraction

Improving the reliability of ML-corrected climate models with novelty detection

Clayton Sanford[1,2], Anna Kwa[1], Oliver Watt-Meyer[1], Spencer K. Clark[1,3], Noah Brenowitz[4], Jeremy McGibbon[1], Christopher Bretherton[1]

[1]Allen Institute for Artificial Intelligence, Seattle, WA
[2]Department of Computer Science, Columbia University, New York, NY
[3]Geophysical Fluid Dynamics Laboratory, NOAA, Princeton, NJ
[4]NVIDIA Corporation, Santa Clara, CA

Key Points:

- A novelty detector is trained to classify whether an atmospheric profile belongs to a high-resolution model's distribution of profiles
- The detector classifies more profiles as novelties in machine-learning corrected simulations that drift further from a reference simulation
- Using the detector to turn off corrections for novel profiles leads to corrected simulations with more consistent and less biased climates

Corresponding author: Anna Kwa, annak@allenai.org

Abstract

The use of machine learning (ML) for the online correction of coarse-resolution atmospheric models has proven effective in reducing biases in near-surface temperature and precipitation rate. However, this often introduces biases in the upper atmosphere and improvements are not always reliable across ML-corrective models trained with different random seeds. Furthermore, ML corrections can feed back on the baseline physics of the atmospheric model and produce profiles that are outside the distribution of samples used in training, leading to low confidence in the predicted corrections. This study introduces the use of a novelty detector to mask the predicted corrections when the atmospheric state is deemed out-of-sample. The novelty detector is trained on profiles of temperature and specific humidity in a semi-supervised fashion using samples from the coarsened reference fine-resolution simulation. Offline, the novelty detector determines more columns to be out-of-sample in simulations which are known, using simple metrics like mean bias, to drift further from the reference simulation. Without novelty detection, corrective ML leads to the development of undesirably large climate biases for some ML random seeds but not others. Novelty detection deems about 21% of columns to be novelties in year-long simulations. The spread in the root mean square error (RMSE) of time-mean spatial patterns of surface temperature and precipitation rate across a random seed ensemble is sharply reduced when using novelty detection. In particular, the random seed with the worst RMSE is improved by up to 60% (depending on the variable) while the best seed maintains its low RMSE.

Plain Language Summary

Corrective machine learning can improve the prediction accuracy of coarse-grid climate models, but also makes them more vulnerable to inputs lying outside the range of training data for the machine learning algorithm. For such inputs, the machine learning may give unreliable results. Using a separate machine learning scheme, we identify out-of-sample data and disable the machine learning correction for these cases. We find that this robustly improves the time-mean temperature and precipitation patterns predicted by ML-corrected climate simulations to be 30-50% better than similar simulations without ML. By improving the accuracy of coarse-grid climate models, this work helps make accurate climate models accessible to researchers without massive computational resources.

1 Introduction

Accurate, reliable climate models are essential for projecting climate change and its impacts. To explore a range of scenarios and account for natural climate variability, climate models must also be computationally quick and affordable. This is typically achieved by using relatively coarse grid resolutions (typically between 50 km and 200 km) and representing processes that operate at finer spatial scales by somewhat empirical human-designed ‘subgrid parameterizations’.

The use of machine learning in atmospheric modeling has taken various forms, including emulating existing physical parameterizations (e.g. Krasnopolsky et al., 2010; Chantry et al., 2021), replacing physics parameterizations by learning from a high-resolution model (e.g. Rasp et al., 2018; Brenowitz & Bretherton, 2019; Yuval & O’Gorman, 2020; Wang et al., 2022), or using ML for online correction of a complete atmospheric model (Watt-Meyer et al., 2021; Bretherton et al., 2022; Clark et al., 2022; Kwa et al., 2022; Chen et al., 2022). Here we will focus on the latter strategy.

Previous works (Rasp et al., 2018; Brenowitz & Bretherton, 2019; Yuval & O’Gorman, 2020; Watt-Meyer et al., 2021) have suggested that correcting or augmenting physics-based climate models with machine learning (ML) can improve weather forecast skill and reduce climate biases. However, ML-augmented models can be susceptible to instabilities (Brenowitz, Beucler, et al., 2020), and their online performance can be sensitive to subtle offline ML training differences, such as random seed selection, in hard-to-predict ways (Brenowitz, Henn, et al., 2020; Wang et al., 2022).

This study draws on the idea of using a compound parameterization (Krasnopolsky et al., 2008; Song et al., 2021) to mask ML models with high uncertainty. Specifically we train a novelty detection algorithm (Hodge & Austin, 2004) and use it to mask ML corrections when the input state is determined to be outside the distribution of the training data. Our approach adds robustness to past approaches (specifically Kwa et al., 2022) while consistently improving temperature and precipitation bias metrics. A preliminary version of this study was presented in Sanford et al. (2022); an important but unrelated software bug fix and some changes in configuration led to substantial changes in interpretation of the effects of the novelty detector, as discussed in Section 2.5.

We model the atmosphere as a discretized system of partial differential equations. The atmospheric state is modeled as $X = (x_1, \ldots, x_N) \in \mathbb{R}^{N \times d}$, a three-dimensional grid of N latitude/longitude coordinates with d-dimensional column vectors concatenating the vertical profiles of gridpoint values of air temperature, specific humidity, winds and other fields. In a 'baseline' model with no added ML corrections, the state of a particular column $x_i \in \mathbb{R}^d$ evolves over time as

$$\frac{dx_i}{dt} = f_i(X, t) \tag{1}$$

for some fixed f_i derived from physically-based assumptions.

The number of grid columns N scales with the inverse square of the desired grid spacing; large N (a fine grid) typically yields more accurate but computationally expensive simulations. While accuracy penalties due to poor grid resolution are expected for small N, coarse-grid simulations are also biased by imperfect representations of subgrid-scale processes like thunderstorms and cloud radiative effects (Zhang & Wang, 2006; Woelfle et al., 2018). ML is an appealing way to de-bias this coarse climate model by predicting and compensating for its error. The ML-corrected model can be written

$$\frac{dx_i}{dt} = f_i(X, t) + g(x_i, \varphi_i; \theta), \tag{2}$$

where $g(\cdot;\theta) : \mathbb{R}^{d+3} \to \mathbb{R}^d$ is a learned function with parameters θ that predicts corrective tendencies from the column, $x_i \in \mathbb{R}^d$, and its insolation, surface elevation, and latitude $\varphi_i \in \mathbb{R}^3$. The ML correction enables the baseline to better approximate a reference fine-grid model while maintaining the underlying physics as the core of the modeling approach (Brenowitz & Bretherton, 2019; Watt-Meyer et al., 2021).

While ML-based models frequently improve overall error, these models—especially deep neural networks—are often not robust, meaning they perform poorly for out-of-sample data. In online application, where predictions are fed back into the model, the corrective ML can induce errors in the overall simulation that accumulate in time, creating large systematic biases and instabilities (Brenowitz, Henn, et al., 2020).

This motivated us to employ semi-supervised novelty detection to predict when a column x_i belongs to the training distribution of g and suppress the tendencies of the ML model if not. This paper shows that strategy can substantially improve the model stability and climate accuracy. With novelty detection, our model has the form

$$\frac{dx_i}{dt} = f_i(X, t) + \eta(x_i; \rho) g(x_i, \varphi_i; \theta), \tag{3}$$

for a novelty detector $\eta(\cdot;\rho) : \mathbb{R}^d \to [0, 1]$ with parameter vector ρ.

2 Methodology

2.1 Dataset

We train the ML tendency correction $g(\cdot;\theta)$ offline as described by Kwa et al. (2022). The training samples $((x_1,\varphi_1),y_1),\ldots,((x_n,\varphi_n),y_n)$ consist of input features and target nudging tendencies y_i(described below) for a set of atmospheric columns sampled at time steps of a nudged coarse model simulation. A neural net with parameters θ is trained to make g best match the nudging tendencies.

The nudged coarse model simulation is constructed to track the evolution of a reference fine-grid no-ML climate model simulation, averaged to the coarse grid cells. Symbolically, the atmospheric state in this reference simulation is denoted $X^{(1)}_{\text{fine}},\ldots,X^{(T)}_{\text{fine}} \in \mathbb{R}^{N\times d}$.

To nudge the coarse simulation to this fine-grid reference, we add a relaxation term to the coarse-grid model of the form

$$y_i := \frac{x_{\text{fine},i} - x_i}{\tau},$$

with a specified nudging timescale $\tau = 3$ hours. By construction, the time-evolving atmospheric state $X^{(1)},\ldots,X^{(T)}$ of this nudged run is approximately (but not exactly) the same as in the fine-grid reference. The y_i are the nudging tendencies that we learn; we denote the $N\times d$ arrays of their values at each time as $Y^{(1)},\ldots,Y^{(T)}$.

For our coarse-grid model f_i, we use a version of NOAA's FV3GFS global weather forecast model (Zhou et al., 2019) with a C48 cubed-sphere grid of approximately 200 km horizontal grid spacing (Putman & Lin, 2007). In this grid, the Earth is divided into 6 square tiles with a 48-by-48 grid imposed on each, for $N = 6\cdot 48^2$ grid columns. This model has 79 vertical levels between the surface and the top of the atmosphere.

The fine-grid reference model (Cheng et al., 2022) is generally similar to the coarse model, with the same number of vertical levels, but has a cubed-sphere C3072 grid with a much finer horizontal grid spacing of approximately 3 km. We used a year of three-hourly reference model output coarsened to the C48 grid by horizontal pressure-level averaging (Bretherton et al., 2022).

Samples are collected from a year-long nudged coarse-grid simulation; the state and nudging tendencies are saved every 3 hours. After dividing this data into interleaved time blocks for the train/test split and randomly subsampling down to 15% of the columns in each timestep, we are left with $n = 2834611$ training samples spanning 2020-01-19 through 2021-01-17.

The same dataset $\mathcal{D}_x = \{x_i \in \mathbb{R}^d : i \in [n]\}$ is used to train the novelty detector $\eta(\cdot;\rho)$. The nudging tendencies y_i are omitted, as the novelty detection procedure requires no labels.

2.2 ML-corrected climate models and data

The novelty detection procedure does not affect the training of the neural nets used to predict the nudging tendencies. We consider two such corrective ML models: g_{Tq} and g_{Tquv}:

- g_{Tq} corrects vertical columns of air temperature T and specific humidity q tendencies, but does not correct winds. That is, x_i is a $d = (2\cdot 79)$-dimensional vector with 79 temperature and 79 humidity coordinates, each corresponding to an atmospheric model level.
- g_{Tquv} also corrects tendencies of the horizontal wind components (u,v) at each level, making x_i a $d = (4\cdot 79)$-dimensional vector.

$g_{\mathrm{Tq}}(\cdot;\theta) : \mathbb{R}^{158} \times \mathbb{R}^3 \to \mathbb{R}^{158}$ predicts the vector of temperature and humidity nudging tendencies y_i from the temperature and humidity profiles x_i, as well as the insolation, surface elevation, and latitude of the corresponding cell (φ_i). Hyperparameters for the corrective ML models were selected after performing a sweep optimized on single-timestep validation loss. We represent $g_{\mathrm{Tq}}(\cdot;\theta)$ as a three-layer dense multi-layer perceptron of width 419. The loss is measured by the mean absolute error (MAE) with L_2 kernel regularization of strength 10^{-4}. We found that models trained with MAE loss were less prone to instabilities and drifts in online simulations than those trained with mean squared error (MSE) loss. We train the model with the Adam optimizer for 500 epochs using a fixed learning rate of 0.00014 and a batch size of 512 samples.

On the other hand, $g_{\mathrm{Tquv}}(\cdot;\theta) : \mathbb{R}^{316} \times \mathbb{R}^3 \to \mathbb{R}^{316}$ is defined as the concatenation of two learned functions for input $x = (x_{\mathrm{Tq}}, x_{\mathrm{uv}}) \in \mathbb{R}^{158} \times \mathbb{R}^{158}$:

$$g_{\mathrm{Tquv}}(x, \varphi; \theta) = (g_{\mathrm{Tq}}(x_{\mathrm{Tq}}, \varphi; \theta_{\mathrm{Tq}}), g_{\mathrm{uv}}(x_{\mathrm{Tq}}, x_{\mathrm{uv}}, \varphi; \theta_{\mathrm{uv}})).$$

$g_{\mathrm{Tq}}(\cdot;\theta_{\mathrm{Tq}})$ is the same as the aforementioned model. $g_{\mathrm{uv}}(\cdot;\theta_{\mathrm{uv}}) : \mathbb{R}^{316} \times \mathbb{R}^3 \to \mathbb{R}^{158}$ is separately trained to infer wind nudging tendencies from temperatures, humidities, and horizontal winds. Besides the different input dimension, $g_{\mathrm{uv}}(\cdot;\theta_{\mathrm{uv}})$ is otherwise structured and trained identically to the other model.

2.2.1 Fixed vertically flipped application of corrective wind tendencies

Kwa et al. (2022) obtained better prognostic simulations using g_{Tq} than by also adding wind tendency correction g_{Tquv}. We have since found this was due to our inadvertently applying the learned wind tendency correction in each column upside-down, such that the correction of the lowest level 79 was applied on the highest level 1, and vice versa. This configuration error, which also affected the wind-corrected simulations discussed by Watt-Meyer et al. (2021), Bretherton et al. (2022) and Clark et al. (2022), arose because FV3GFS uses opposite vertical indexing of grid levels in the physical parameterizations and dynamical core. After fixing this error, including corrective wind tendencies no longer leads to numerical instability, and most metrics of 3-7 day weather skill (e.g. RMSEs of 850hPa temperature) are significantly improved.

We consider both corrective approaches to trace through the effects of rectifying this error on our results, and to show that the novelty detection is useful for both correction methods.

2.3 Novelty detection

The novelty detector η predicts whether a column x belongs within the support of the training set, and we use it to mask the ML-predicted corrections. Specifically, if a column is determined to not be a novelty, then we let $\eta(x;\rho) = 1$ (recall Equation 3) to take full advantage of the learned correction $g(x, \varphi; \theta)$; otherwise, we ignore $g(x, \varphi; \theta)$ by setting $\eta(x;\rho) = 0$.

Novelty detection is a well-studied semi-supervised learning problem about estimating the support of a dataset using only positive examples (Hodge & Austin, 2004). We frame the problem as novelty detection rather than outlier detection (an unsupervised problem with mixture of in-distribution and out-of-distribution samples) or standard two-class supervised classification, because we have no dataset of representative out-of-distribution samples and constructing such a dataset would introduce additional model-dependence into this process.

There are many known approaches to novelty detection, including local-outlier factor (Breunig et al., 2000), k-means clustering (Nairac et al., 1999), and minimum-volume ellipsoid estimation (Van Aelst & Rousseeuw, 2009). Our exploratory work considers two approaches: a simple "min-max" novelty detector and a one-class support vector machine (OCSVM). For each of these we consider novelty detectors η_{T} with 79-dimensional temperature vectors as input and η_{Tq} with 158-dimensional combined temperature and specific humidity vectors.

We did not consider novelty detectors with wind inputs. Adding more inputs to the OCSVM classifier requires further hyperparameter tuning (see Appendix A) to keep the evaluation time low enough to be usable within prognostic simulations; we therefore limit the scope of this work to out-of-sample detection on temperature and specific humidity fields.

Naive "min-max" novelty detector The min-max novelty detector considers the smallest axis-aligned hyper-rectangle that contains all training samples and categorizes any sample outside the rectangle as a novelty:

$$\eta_{\mathrm{minmax}}(x; (x_{\min}, x_{\max})) = \begin{cases} 1 & \text{if } x_k \in [x_{\min,k}, x_{\max,k}] \; \forall k \in [d], \\ 0 & \text{otherwise,} \end{cases}$$

for $x_{\min,k} = \min_{i,t} x_{i,k}^{(t)}$ and $x_{\max,k} = \max_{i,t} x_{i,k}^{(t)}$ as the minimum and maximum over the training data of the kth feature. While efficient, this novelty detector cannot identify irregular correlations between input features that nevertheless lie within the bounding box.

One-class support vector machine (OCSVM) The one-class SVM algorithm of Schölkopf et al. (2001) repurposes the SVM classification algorithm to estimate the support of a distribution by finding the maximum-margin hyperplane separating training samples from the origin. The OCSVM has been applied to novelty detection for genomics (Sommer et al., 2017), video footage (Amraee et al., 2018), propulsion systems (Tan et al., 2019), and the internet of things (Yang et al., 2021).

We normalize each input x_i and utilize the kernel trick, lifting it to the infinite-dimensional feature space $\phi(x_i)$ corresponding to the radial basis function (RBF) kernel $\kappa_\gamma(x, x') = \exp(-\gamma \|x - x'\|_2^2)$. We parameterize the novelty detector with $\rho = (\alpha, \xi, \gamma)$ in its dual form,

$$\eta_{\mathrm{OCSVM}}(x; (\alpha, \xi, \gamma)) = \begin{cases} 1 & \text{if } \sum_{i=1}^{n} \alpha_i \kappa_\gamma(x, x_i) \geq \xi, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

The sensitivity of the novelty detector can be adjusted by choosing a cutoff $\xi > 0$. The learnable real-valued weights $\alpha_i \geq 0$ correspond to each training sample x_i. Intuitively, a large α_i indicates that the proximity of x to x_i indicates that x is likely in the support; a small (or zero-valued) α_i means that the novelty of samples can be determined without measuring their proximity to x_i. The goal is to find a relatively small subsample of training samples x_i and corresponding nonzero weights α_i that can be used to confidently and efficiently assess whether x is out of sample.

The weights are learned by solving a quadratic program based on the training data. The number of nonzero α_i depends on γ and a regularization parameter ν. The prediction rule depends exclusively on the support vectors, or the training samples x_i with $\alpha_i > 0$. To obtain a robust and computationally efficient novelty detector, for a given γ we choose ν to ensure the number of support vectors is on the order of at most 10^4, less than 0.5% of the training data sample. Smaller values of γ correspond to novelty detectors with highly smoothed support estimations that may be larger than necessary, while large γ provides a smaller and perhaps more topologically complex region.

We explain the process of OCSVM parameter selection and the resulting trade-offs more comprehensively in Appendix A and Section 4. For the main results presented in 3, we chose $\gamma = 4/79$, $\nu = 10^{-4}$, and $\xi = 0.12$, for which the novelty detector classifies none of the training data and an acceptably small 2% of a withheld test set of the reference data as being out-of-sample.

2.4 Computing scalar metrics

We measure the success of a coarse-grid simulated run by computing the root mean square error (RMSE) of time-averaged quantities (850hPa and 200hPa temperature, surface precipitation, total precipitable water) with respect to those same quantities for the coarsened fine-grid run. We compute the RMSE of the time-averaged field s as follows:

$$\mathrm{RMSE}(s) = \sqrt{\sum_{i=1}^{N} a_i \left(\frac{1}{T} \sum_{t=1}^{T} \left(\hat{s}_i^{(t)} - s_{\mathrm{fine},i}^{(t)} \right) \right)^2}, \tag{5}$$

where $\hat{s}_i^{(t)}$ and $s_{\mathrm{fine},i}^{(t)}$ denote the field value at grid cell $i \in [N]$ and time $t \in [T]$ in our coarse-grid and the reference fine-grid simulations respectively, and a_i are the normalized area weights of grid cells.

2.5 Methodological updates vs. Sanford et al. (2022)

We made two important methodological updates in this study compared to a similar recent work on which it is based (Sanford et al., 2022). Firstly, we fixed the previous error (see Section 2.2.1), discovered after that earlier work, where the ML wind tendencies in each grid column were applied with inverted vertical indexing during online simulations. The second change is related to the application of the ML corrections g_{Tq} and g_{Tquv} in the upper atmosphere. Sanford et al. (2022) followed the approach of Kwa et al. (2022), in which the ML-predicted tendencies in the top three model layers were not applied as corrections. The rationale was that the sponge layer differences between low and high resolution models was a process we did not wish to correct, and there were relatively large magnitude nudging tendencies at these levels. In this study, we use a more aggressive tapering in which the ML-predicted outputs are tapered to zero throughout the uppermost 25 model levels using an exponential decay, as in Equation 6 of Clark et al. (2022). This improves the simulation of lower atmospheric air temperatures, and more importantly, helps prevent large upper atmospheric temperature drifts when using ML corrections of horizontal winds. Both of these changes improve the ML-corrected simulations described by Equation 2 and impose a higher bar for the novelty detection to add value.

3 Results

3.1 Offline application of novelty detection

Before integrating a novelty detector into online simulations with an ML-corrected climate model, we test it offline on data produced by the preexisting simulations. We compare the frequency of offline novelty detection for datasets generated from the first 16 weeks of three C48 simulations – a no-ML baseline model simulation and two g_{Tquv}-corrected simulations that differ only in the random initial seed used in training the g_{Tquv} models. The g_{Tquv} seed 0 run has the largest yearly-mean precipitable water RMSE (4.4 kg/m^2) across a set of four g_{Tquv} simulations, while the seed 3 run has the smallest (2.4 kg/m^2), slightly smaller than that of the baseline run (2.7 kg/m^2). Results with these two seeds demonstrate the difference in out-of-sample fraction between 'high error' and 'low error' ML-corrected simulations. Feedback loops between less reliable ML corrections and out-of-sample column states may exacerbate mean-state drifts, showing up as locally higher offline novelty fractions. The baseline simulation tests the extent to which mean-state biases developing in a conventional climate model lead to detectable novelties.

Figure 1 focuses on the first 16 simulated weeks of the simulation to make the drifts into out-of-sample states more visible. Within a few days, the baseline model moistens relative to the reference model until it generates enough clouds and precipitation to bal-

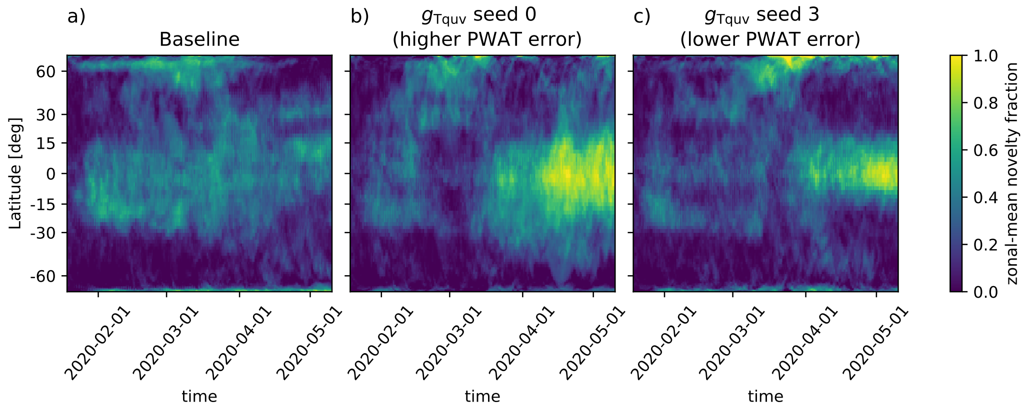


Figure 1. Zonal-mean fraction of novelties detected by the $\eta_{\mathrm{Tq,OCSVM}}$ novelty detector over the first 16 weeks of the (a) baseline, (b) seed 0 g_{Tquv} , and (c) seed 3 g_{Tquv} simulations.

ance surface evaporation, after which it settles into a new, slightly biased equilibrium in which about 25% of the columns are flagged as novelties.

Initially, the seed 0 and seed 3 g_{Tquv} corrections both have the intended effect of keeping the global state closer to the fine-grid reference distribution. These ML-corrected g_{Tquv} runs have lower global novelty fractions than the baseline over the first two months, particularly in the tropics. However, from March onward, the novelty fraction in the baseline tropics plateaus, while both g_{Tquv} simulations continue to drift farther out-of-sample in the tropics.

By the end of the 16 weeks shown in Figure 1, the "high error" seed 0 g_{Tquv} simulation has roughly twice as many out-of-sample columns compared to the baseline and "low error" seed 3 runs. This demonstrates that suboptimal ML corrections (as in the seed 0 g_{Tquv} model) can indeed push the state further out of the training set distribution, setting the stage for less reliable ML corrections that further exacerbate climate drifts.

3.2 Online novelty detection improves temperature and precipitation predictions

We assess the utility of the novelty detectors by incorporating $\eta(\cdot; \rho)$ into the coarse grid model and numerically simulating equation (3) for one year. We compare the predicted atmospheric states $\hat{x}_i$ to $x_{\mathrm{fine},i}$ using the RMSE of four time-averaged diagnostics calculated using equation (5): air temperatures at pressures of 200 hPa and 850 hPa (T200, T850) representative of the lower and upper troposphere, surface precipitation rate (SP)[1], and precipitable water (PWAT)[2].

Table 1 compares the performance of six global simulations. The first is the no-ML baseline simulation; the next two are ML-corrected runs without and with wind tendency corrections; and the remaining three simulations use g_{Tquv} corrections and include novelty detection from equation (3) – these differ in the choice of novelty detector η and its inputs. The η_{Tq} OCSVM uses the same parameter choices as for the offline comparisons.

[1] Current climate models make less consistent predictions of regional shifts in precipitation than of surface temperatures; contrast sections B.2.1 and B.3.1 of IPCC (2021).

[2] PWAT is the total mass of water contained in a vertical atmospheric column per cross-sectional area and is highly correlated with the regional precipitation rate (Bretherton et al., 2004).

Table 1. The RMSE scores of time-averaged metrics and novelty detection rates for year-long simulations. Values for ML-corrected runs are the mean, with standard deviation in parentheses, across the four random seeds. The "% Novelty" column represents the percent of columns over the simulated year which were classified as out-of-sample and did not receive ML corrections. Metrics are 200- and 850-hPa temperature (T200, T850), surface precipitation rate (SP) and precipitable water (PWAT). For each metric, the run with the lowest RMSE is bolded.

Run	% Novelty	T200 (K)	T850 (K)	SP (mm/day)	PWAT (kg/m^2)
Baseline	-	2.48	2.09	1.78	2.79
g_{Tq}	-	2.50 (0.40)	1.97 (0.08)	1.52 (0.07)	3.97 (0.29)
g_{Tquv}	-	3.30 (0.49)	1.31 (0.14)	1.40 (0.12)	3.40 (0.73)
$g_{\mathrm{Tquv}}, \eta_{\mathrm{T,minmax}}$	0.6 (0.3)	3.04 (0.65)	**1.29 (0.06)**	1.36 (0.07)	3.28 (0.72)
$g_{\mathrm{Tquv}}, \eta_{\mathrm{T,OCSVM}}$	5.0 (1.0)	2.84 (0.49)	1.38 (0.09)	1.37 (0.08)	3.36 (0.83)
$g_{\mathrm{Tquv}}, \eta_{\mathrm{Tq,OCSVM}}$	20.6 (4.8)	**1.24 (0.05)**	1.30 (0.08)	**1.29 (0.07)**	**2.38 (0.37)**

For the η_{T} OCSVM, which uses fewer inputs, we use the same $\gamma = 4/79$ and $\nu = 10^{-4}$ but readjust the cutoff ξ to 2.02 to the minimum needed to suppress T-only novelties within the training dataset. For all the configurations except the baseline, we perform an ensemble of simulations using four identically-trained ML-correction models g initialized with different random seeds.

Without a novelty detector, the conclusions for the g_{Tq} model (ML-corrected temperature and humidity tendencies only) are similar to Kwa et al. (2022). The metrics (second row in Table 1) are 10-20% better than for the baseline model, except for the PWAT RMSE which worsens. Adding corrective ML for winds (third row in Table 1) significantly improves the 850 hPa air temperature errors (ensemble-mean RMSE decreases from 1.97 K to 1.31 K), somewhat improves SP and PWAT, but substantially worsens the T200 RMSE.

The min-max novelty detector (fourth row in Table 1) slightly improves the RMSEs but has limited impact since it activates only rarely (in 0.6% of atmospheric columns, as shown in the second column of the table). This indicates the importance of bounding the data distribution more tightly than a high-dimensional box. The $\eta_{\mathrm{T,OCSVM}}$ novelty detector classifies a higher fraction of columns as novelties (5%) than the min-max detector, but the overall RMSE for the $g_{\mathrm{Tquv}}, \eta_{\mathrm{T,OCSVM}}$ simulations are mostly on par with the g_{Tquv} results without novelty detection, with the exception of further improvements in T200 RMSE.

The $\eta_{\mathrm{Tq,OCSVM}}$ novelty detector, on the other hand, improves 200 hPa air temperature, surface precipitation, and precipitable water RMSEs by 62%, 8%, and 30% respectively, compared to the g_{Tquv} simulations without novelty detection. To achieve these improvements, the OCSVM novelty detectors activate in 21% of all atmospheric columns, averaged over the course of the year-long simulations. If compared to the same 16 week time period as the offline analysis of g_{Tquv} runs without novelty detection in Section 3.1, online novelty detection reduces the novelty fraction in ML-corrected runs by roughly half. In summary, suppressing ML corrections to columns with atypical temperature and specific humidity profiles helps keep the g_{Tquv}-corrected model within the envelope of its training data, where it is skillful in reducing temperature and humidity biases.

Figure 2 shows the RMSE of time-mean surface precipitation, 200 hPa and 850 hPa temperature, and precipitable water across individual ensemble members of simulations using g_{Tq}, g_{Tquv}, and $g_{\mathrm{Tquv}}, \eta_{\mathrm{Tq,OCSVM}}$. This illustrates that the $\eta_{\mathrm{Tq,OCSVM}}$ novelty detection substantially reduces the variance in skill across the ML-corrected runs (also demon-

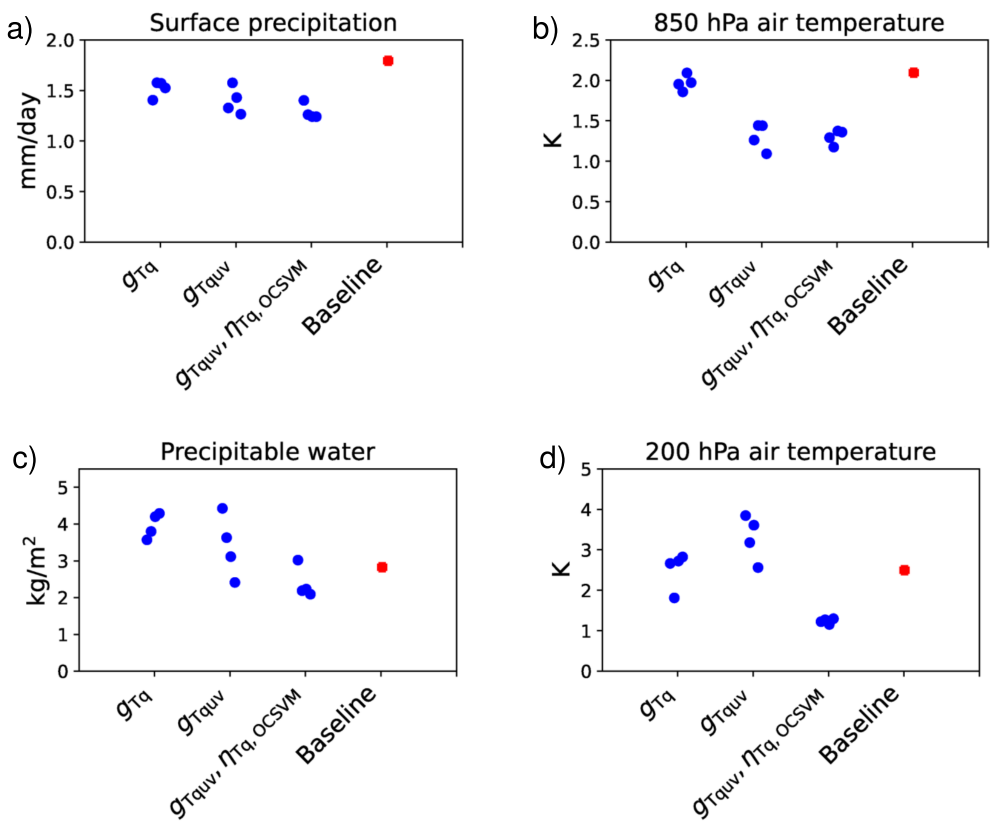


Figure 2. RMSE of time-mean fields in groups of ML-corrected simulations and the baseline prognostic run. Each group of four blue points shows a range of results across four randomly seeded corrective-ML models. The same randomly seeded g_{Tq} models are used in all ML-corrected groups. The same four g_{uv} models are used in both the g_{Tquv} and $g_{\mathrm{Tquv}}, \eta_{\mathrm{Tq,OCSVM}}$ groups.

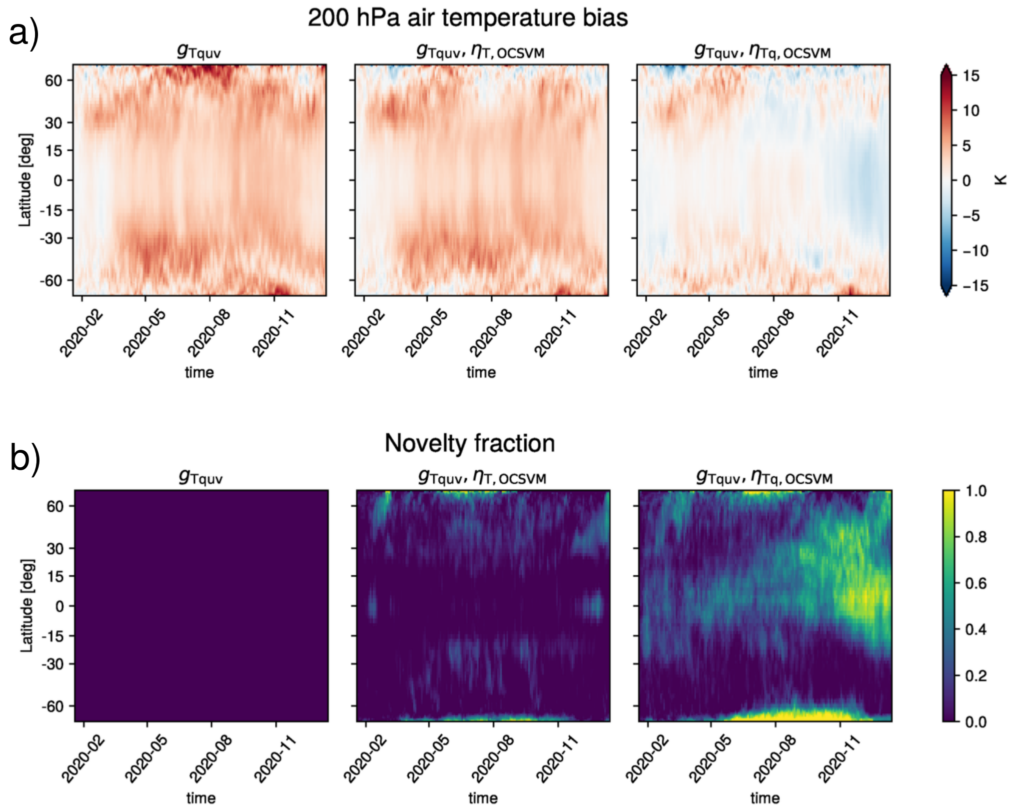


Figure 3. Time versus zonal-mean plots visualizing upper-atmospheric temperature biases (against the fine-grid reference simulation) at the 200 hPa pressure level (top) and fractions of novelties identified (bottom) by three different models initialized from random seed 0 (left to right): (1) the ML-corrected climate model g_{Tquv} without novelty detection, (2) g_{Tquv} with OCSVM novelty detection $\eta_{\mathrm{T,OCSVM}}$ using temperature as the input feature, and (3) g_{Tquv} with OCSVM novelty detection $\eta_{\mathrm{Tq,OCSVM}}$ using temperature and specific humidity as input features.

strated by the standard deviations reported in parentheses in Table 1), especially for precipitable water and 200 hPa temperature. The novelty detection reduces variance and improves the overall ensemble skill by bringing the worst-performing g_{Tquv} seeds closer in line with the better performers.

3.3 Improvements for a particular ML-corrected simulation

In this subsection, the ML-corrected simulation results are shown just for the worst g_{Tquv} seed (0), to provide a clear illustration of how novelty detection especially benefits poorly-performing prognostic runs. This seed's $g_{\mathrm{Tquv}}, \eta_{\mathrm{Tq,OCSVM}}$ simulation had a novelty fraction of 24.3%, slightly higher than the ensemble mean of 20.6%.

3.3.1 Zonal-mean biases

Figure 3 compares the time evolution of zonal-mean 200 hPa air temperature biases in three ML-corrected year-long simulations: g_{Tquv} without novelty detection, and

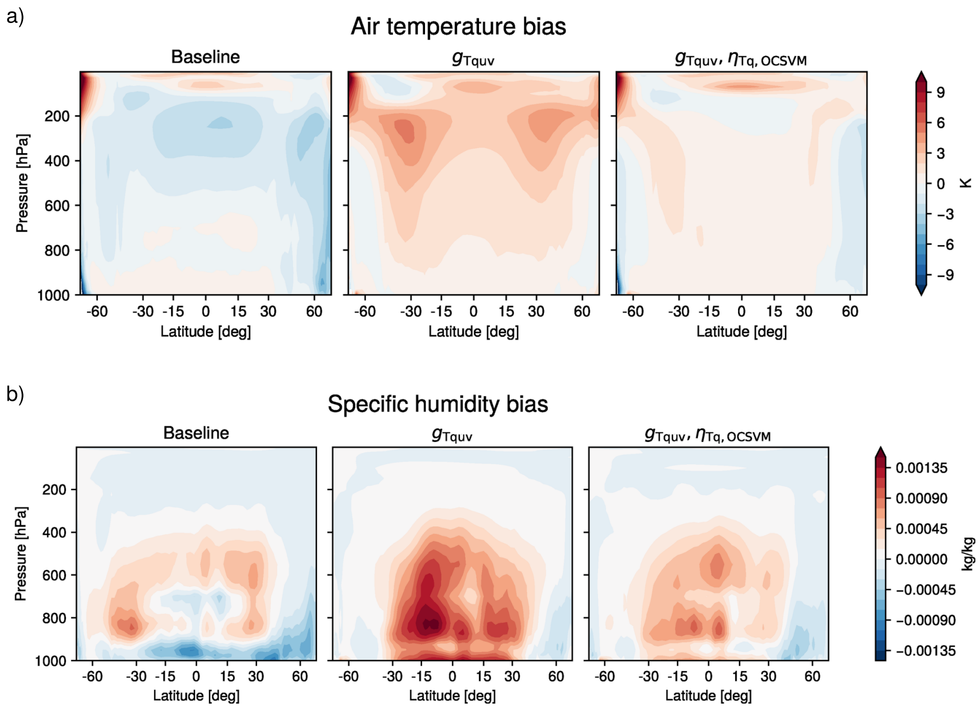


Figure 4. Annual-averaged zonal mean temperature (top) and humidity (bottom) biases plotted over pressure levels, for the baseline model (left) and seed-0 g_{Tquv} models with no novelty detection (center) and with $\eta_{\mathrm{Tq,OCSVM}}$ novelty detection (right).

two simulations with novelty detectors $\eta_{\mathrm{T,OCSVM}}$ and $\eta_{\mathrm{Tq,OCSVM}}$ that use different feature sets.

The ML-corrected g_{Tquv} model without novelty detection develops a significant 5-10 K warm bias in 200 hPa air temperature across latitudes. The temperature-only novelty detection in $g_{\mathrm{Tquv}}, \eta_{\mathrm{T,OCSVM}}$ removes the the largest magnitude warm bias at the North Pole during boreal summer, but otherwise does not prevent the global warm drift. Though the $g_{\mathrm{Tquv}}, \eta_{\mathrm{T,OCSVM}}$ simulation develops 5-10 K biases within the first 16 weeks, the $\eta_{\mathrm{T,OCSVM}}$ OCSVM activates infrequently as it still classifies these columns' temperature profiles as lying within the training distribution, presumably due to the large weather-associated variability of temperature sampled therein.

The prognostic run in the right column of Fig. 3 shows that using specific humidity inputs in addition to temperature inputs is necessary for successful bias reduction via novelty detection. This greatly increases the rate of out-of-sample classification, especially in the tropics. The 200 hPa temperature bias is dramatically reduced out to high latitudes, despite the majority of the novelty detection occurring in the tropics. We speculate that this is due to changes in tropical convection, where the $\eta_{\mathrm{Tq,OCSVM}}$ novelty detector is most active other than extreme polar latitudes.

Figure 4 shows sections of time- and zonal-mean air temperature and specific humidity biases. Instead of the $g_{\mathrm{Tquv}}, \eta_{\mathrm{T,OCSVM}}$ run, Figure 4 includes a baseline (no-ML) simulation for comparison, since that is what we are aiming to improve on. The baseline model air temperature is biased low in the tropical stratosphere and throughout the column in high northern latitudes. The ML-corrected g_{Tquv} model without novelty detection corrects the cold bias at high northern latitudes but develops an overall warm

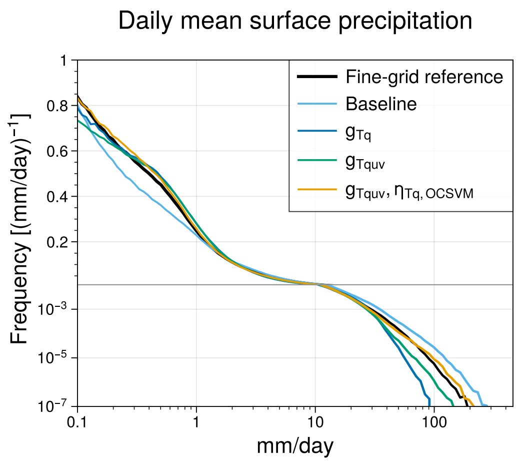


Figure 5. Probability distribution function of daily-mean precipitation from all grid columns around the globe, shown for the fine-grid reference, baseline, ML-corrected g_{Tquv} run without novelty detection, and ML-corrected $g_{\mathrm{Tquv}}, \eta_{\mathrm{Tq,OCSVM}}$ run. The y-axis uses linear scaling above 0.01 $(\mathrm{mm/day})^{-1}$ and log scaling below.

bias that is largest in the extratropical stratosphere. Adding the $\eta_{\mathrm{Tq,OCSVM}}$ novelty detector on top of the g_{Tquv} corrections removes most of this stratospheric warm bias.

Similarly, the ML-corrected g_{Tquv} model without novelty detection develops a tropical moist bias in specific humidity that is larger in magnitude than the baseline biases in both the boundary layer and the troposphere. Adding the $\eta_{\mathrm{Tq,OCSVM}}$ novelty detector greatly reduces this bias.

3.4 Daily-mean precipitation distribution

The ML-corrected $g_{\mathrm{Tquv}}, \eta_{\mathrm{Tq,OCSVM}}$ simulation also captures the global-mean probability distribution function (pdf) of daily mean precipitation in the reference fine-grid simulation better than the baseline (no-ML) and g_{Tquv} approaches (Figure 5). The baseline run underestimates the frequency of low daily-mean precipitation below a few mm/day, while the ML-corrected simulations more closely match the fine-grid reference at the low end of the distribution. The baseline run over-estimates the high-precipitation tail of the target pdf, while the g_{Tquv} run underestimates the pdf in the tail. The $g_{\mathrm{Tquv}}, \eta_{\mathrm{Tq,OCSVM}}$ run matches the tail of the global precipitation pdf more closely up to rates over 100 mm/day.

4 Varying novelty detector sensitivity

Section 3 considered an OCSVM with $\gamma = 4/79$ and cutoff ξ set to the maximum score observed in the training data[3]. This model, whether applied only to temperature or to both temperature and humidity, appears to find a consistent "sweet spot" between the baseline run and the ML-corrected run with no novelty detection that reduces the mean-state drifts of both approaches. This section presents a sensitivity study that supports this finding by considering several choices of γ and varying ξ to adjust the aggressiveness of the novelty detector. We show that these approaches interpolate between the

[3] See Appendix A for a more thorough discussion of how the value of cutoff ξ impacts novelty frequency in online simulations.

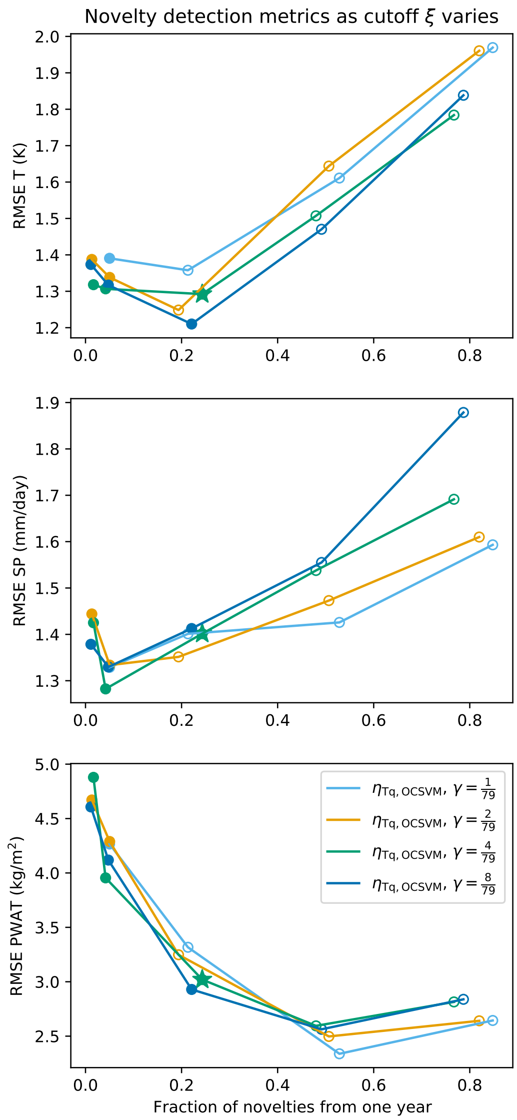


Figure 6. RMSE of time-averaged 850hPa temperature (top), surface precipitation (center), and precipitable water (bottom) of year-long global C48 simulations, all with ML-correction g_{Tquv} and novelty detector $\eta_{\mathrm{Tq,OCSVM}}$ with kernel inverse radius $\gamma \in \{\frac{1}{79}, \frac{2}{79}, \frac{4}{79}, \frac{8}{79}\}$. The plots show each error metric as a function of the total fraction of identified novelties (a monotonic increasing function of ξ) on the x-axis. The choices of ξ are given in Table A1 in Appendix A. The single green star marker represents the OCSVM parameters used in the Results section. Filled markers indicate consistent novelty detectors that classify no more than 5% of the holdout reference training dataset as out-of-sample; open circles indicate inconsistent detectors.

baseline and ML-corrected run as the cutoffs change, and that choosing an intermediate model that categorizes a substantial fraction of samples as novelties balances the trade-off between the quality of temperature and surface precipitation estimates and of precipitable water estimates.

In Figure 6, we consider an ML-corrected model g_{Tquv} augmented with an OCSVM novelty detector $\eta_{\mathrm{Tq,OCSVM}}$ with various choices of inverse radius parameter γ and cutoff parameter ξ. We plot the error metrics as a function of the fraction of novelties identified online for each cutoff. We find that an intermediate cutoff balances strong performance on temperature and surface precipitation (for which the ML-correction-only simulation has a lower RMSE than the baseline simulation) and total precipitable water estimates (which are better predicted by the baseline model than the ML-correction-only simulation). Optimal temperature and precipitable water predictions generally occur when approximately 20% and 60% of samples are categorized as novelties, respectively (and hence suppressed). The plots demonstrate that this approach effectively interpolates between those two extreme cases and that the cutoff ξ used in the preceding section lies near that sweet spot. The figure also demonstrates that different combinations of radius parameter γ and cutoff ξ result in similarly performing simulations when the fraction of novelties detected is the same.

As the classification cutoff ξ is increased, a greater fraction of samples from the reference dataset distribution will also be classified as out-of-sample. Detectors are considered inconsistent if they return a significant novelty fraction when evaluated on a holdout set from the training data, as this means that the detector has a higher false positive rate in flagging samples as novelties when they are still within the training distribution. In this analysis we set a false-positive threshold of 5% to determine which (γ, ξ) combinations are consistent OCSVMs. OCSVMs which classify $> 5\%$ of the holdout reference data as out-of-sample are deemed inconsistent and indicated as open circles in Figure 6. These include all detectors that classify less 75% of online ML-corrected samples as typical and, for certain γ, even detectors classifying up to 95% as such. That is, the best climate performance using this ML correction model is found by using the maximum ξ consistent with the false-positive threshold on the withheld reference data.

5 Conclusion and future work

This study demonstrates that applying novelty detection to ML-corrected coarse-grid atmospheric climate models can improve the quality and reliability of their temperature and precipitation estimates. Offline, a novelty detection algorithm trained on samples from a coarsened high-resolution simulation tends to classify more columns as novelties in runs that drift further from the high-resolution reference. When applied online to mask ML-predicted corrective tendencies, the novelty detector maintains or improves the spatial patterns of time-mean surface precipitation rate, lower and upper atmospheric temperature and precipitable water. Furthermore, for an ensemble of ML-corrected simulations (in which each simulation uses an ML model trained with a different random seed initialization of weights), use of novelty detection decreases the spread in model skill across the ensemble. This is a valuable property, since online use of ML parameterizations can be highly sensitive to subtle changes in the offline training, such as random seed (e.g. Wang et al., 2022).

Future work can build on this effort by experimenting with different novelty detection approaches, OCSVM kernels, inputs to η, and methods for integrating the novelty detector into the ML-corrected climate model. Practical implementation of the novelty detector can become a simulation bottleneck if the number of support vectors (Appendix A) is too high. For the settings used in Section 3, the novelty detector roughly doubled the wall clock time per simulation timestep. It would be worth further investigation into how few support vectors are needed to improve ML-corrected simulations online. In addition the more classical ML approaches to novelty detection explored here, future work may consider using neural networks directly for density estimation for the purpose of novelty detection. Finally, further analysis of the character of the out-of-sample behaviors that are being detected by the trained novelty detectors could help us better understand their causes.

Appendix A Parameterization of the One-Class Support Vector Machine

As described in Section 2.3, we use a one-class SVM to predict whether a 158-dimensional column of temperatures and humidities is out-of-distribution and hence, whether we should suppress the learned ML-corrections. Implementation details and choices are discussed below.

Given a set of in-distribution training data $x_1, \dots, x_n \in \mathbb{R}^{158}$, we learn a decision rule that categorizes a new training sample x as in-distribution or not depending based on the correspondence with each training sample x_i according to kernel function κ_γ:

$$\eta_{\mathrm{OCSVM}}(x; (\alpha, \xi, \gamma)) = \mathbb{1}\left\{\sum_{i=1}^{n} \alpha_i \kappa_\gamma(x, x_i) \geq \xi\right\}.$$

The model includes learned weights $\alpha \in [0,1]^n$, a fixed kernel inverse-radius parameter γ, and a cutoff parameter ξ.

We use the radial basis function (RBF) kernel $\kappa_\gamma(x, x') = \exp(-\gamma \|x - x'\|_2^2)$ because of its expressibility and due to the ease of trading off bias and variance with its inverse-radius parameter γ. A large choice of γ ensures that $\kappa_\gamma(x, x')$ only has non-negligible output if x is extremely close to x', while smaller γ selections cause a large "ball" of x around x' to all have $\kappa_\gamma(x, x') \approx 1$. Choosing large γ makes for a more expressive classifier that can be used to fit any training data, but raises the risk of classifying many 'holes' in between training data samples as out-of-sample. A smaller γ imposes a smoothing effect on the learned classifier. The default SVM setting in scikit-learn is $\gamma = \frac{1}{\#\text{ features}} = \frac{1}{2\cdot 79}$. For our application, we find that a larger choice of γ tends to produce better outcomes and focus our study on four choices: $\gamma \in \{\frac{1}{79}, \frac{2}{79}, \frac{4}{79}, \frac{8}{79}\}$.

The scikit-learn implementation of an OCSVM uses a regularization parameter ν in the training procedure to trade off classification accuracy and model simplicity when learning weights $\alpha \in [0,1]^n$ (Schölkopf et al., 2000). ν does so by regulating the number of allowable support vectors, which are samples x_i that have respective weight $\alpha_i > 0$, which in turn scales the computational cost of each application of the OCSVM. Choosing a large value of ν puts a greater premium on categorizing every sample correctly by using more support vectors. Here, we use a parameter search to choose a ν for each γ that results in roughly 10^4 support vectors.

Finally, the cutoff ξ affects the sensitivity of the learned novelty detector. A large choice of ξ causes an aggressive detector that categorizes a large number of samples as novelties (and hence, turns of the ML-corrected tendencies frequency), while a small ξ classifies more samples as in-distribution. We calibrate the sensitivity by drawing samples from a full year of an ML-corrected run and choosing a cutoff ξ_p such that a fraction p of the given data are categorized as in-distribution; a larger choice of p results in a smaller ξ_p. For the sensitivity study in Section 4, we consider the corresponding ξ_p choices for each γ for $p \in \{0.25, 0.5, 0.75, 0.95, 0.99\}$. In Table A1, we give the respective choices of ν and ξ_p for each γ.

Table A1. One-class SVM parameterizations. For each kernel radius γ, we select a regularization parameter ν in order to constrain the number of support vectors to roughly 10000 for computational efficiency, which is in turn used to train a parameter vector α. Five cutoffs ξ are identified to adjust the conservatism of the model: ξ_p is chosen to ensure that a p fraction of the training dataset is categorized as in-distribution, i.e. $\eta_{\mathrm{OCSVM}}(x; (\alpha, \xi_p, \gamma)) = 1$.

γ	ν	# SVs	$\xi_{0.25}$	$\xi_{0.5}$	$\xi_{0.75}$	$\xi_{0.95}$	$\xi_{0.99}$
$\frac{1}{79}$	$5 \cdot 10^{-3}$	14365	351	321	289	227	153
$\frac{2}{79}$	$5 \cdot 10^{-3}$	15029	80	70	60	42	22
$\frac{4}{79}$	$1 \cdot 10^{-4}$	16030	0.18	0.15	0.12	0.065	0.023
$\frac{8}{79}$	$4 \cdot 10^{-6}$	12152	$5.9 \cdot 10^{-4}$	$4.4 \cdot 10^{-4}$	$2.8 \cdot 10^{-4}$	$9.3 \cdot 10^{-5}$	$1.7 \cdot 10^{-5}$

Open Research

Code used to run these experiments is available at the Github repositories `https://github.com/ai2cm/out-of-sample` and `https://github.com/ai2cm/fv3net`, which are archived at `https://zenodo.org/record/7872723` and `https://zenodo.org/record/7872718`, respectively. The coarsened fine-grid data used for initial conditions and in the nudged coarse-grid simulation is available upon request through a Google Cloud Storage 'requester pays' bucket.

Acknowledgments

This work was started when Sanford was a summer intern with the AI2 Climate Modeling group. AI2 is supported by the estate of Paul G. Allen. We appreciate helpful conversations with Daniel Hsu about framing the problem as novelty detection. We thank NOAA-GFDL for running the 1-year X-SHiELD simulation using the Gaea computing system, which was used as the reference fine-grid dataset in this work. We also acknowledge NOAA-GFDL, NOAA-EMC, and the UFS community for sharing code, forcing data, and software packages.

References

Amraee, S., Vafaei, A., Jamshidi, K., & Adibi, P. (2018). Abnormal event detection in crowded scenes using one-class svm. *Signal, Image and Video Processing*, *12*(6), 1115–1123. Retrieved from `https://doi.org/10.1007/s11760-018-1267-z` doi: 10.1007/s11760-018-1267-z

Brenowitz, N. D., Beucler, T., Pritchard, M., & Bretherton, C. S. (2020). Interpreting and stabilizing machine-learning parametrizations of convection. *Journal of the Atmospheric Sciences*, *77*(12), 4357 - 4375. Retrieved from `https://journals.ametsoc.org/view/journals/atsc/77/12/jas-d-20-0082.1.xml` doi: 10.1175/JAS-D-20-0082.1

Brenowitz, N. D., & Bretherton, C. S. (2019). Spatially extended tests of a neural network parametrization trained by coarse-graining. *Journal of Advances in Modeling Earth Systems*, *11*(8), 2728-2744. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019MS001711` doi: https://doi.org/10.1029/2019MS001711

Brenowitz, N. D., Henn, B., Clark, S., Kwa, A., McGibbon, J., Perkins, W. A., ... Bretherton, C. S. (2020). Machine learning climate model dynamics: Offline versus online performance. In *Neurips 2020 workshop on tackling climate change with machine learning.* Retrieved from `https://`

www.climatechange.ai/papers/neurips2020/50

Bretherton, C. S., Henn, B., Kwa, A., Brenowitz, N. D., Watt-Meyer, O., McGibbon, J., . . . Harris, L. (2022). Correcting coarse-grid weather and climate models by machine learning from global storm-resolving simulations. *Journal of Advances in Modeling Earth Systems*, *14*(2), e2021MS002794. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021MS002794 (e2021MS002794 2021MS002794) doi: https://doi.org/10.1029/2021MS002794

Bretherton, C. S., Peters, M. E., & Back, L. E. (2004). Relationships between water vapor path and precipitation over the tropical oceans. *Journal of Climate*, *17*(7), 1517 - 1528. Retrieved from https://journals.ametsoc.org/view/journals/clim/17/7/1520-0442_2004_017_1517_rbwvpa_2.0.co_2.xml doi: 10.1175/1520-0442(2004)017⟨1517:RBWVPA⟩2.0.CO;2

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). Lof: Identifying density-based local outliers. In *Proceedings of the 2000 acm sigmod international conference on management of data* (p. 93–104). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/342009.335388 doi: 10.1145/342009.335388

Chantry, M., Hatfield, S., Dueben, P., Polichtchouk, I., & Palmer, T. (2021). Machine learning emulation of gravity wave drag in numerical weather forecasting. *Journal of Advances in Modeling Earth Systems*, *13*(7). Retrieved from https://doi.org/10.1029/2021ms002477 doi: 10.1029/2021ms002477

Chen, T.-C., Penny, S. G., Whitaker, J. S., Frolov, S., Pincus, R., & Tulich, S. (2022). Correcting systematic and state-dependent errors in the noaa fv3-gfs using neural networks. *Journal of Advances in Modeling Earth Systems*, *14*(11), e2022MS003309. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2022MS003309 (e2022MS003309 2022MS003309) doi: https://doi.org/10.1029/2022MS003309

Cheng, K.-Y., Harris, L., Bretherton, C., Merlis, T. M., Bolot, M., Zhou, L., . . . Fueglistaler, S. (2022). Impact of warmer sea surface temperature on the global pattern of intense convection: Insights from a global storm resolving model. *Geophysical Research Letters*, *n/a*(n/a), e2022GL099796. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2022GL099796 (e2022GL099796 2022GL099796) doi: https://doi.org/10.1029/2022GL099796

Clark, S. K., Brenowitz, N. D., Henn, B., Kwa, A., McGibbon, J., Perkins, W. A., . . . Harris, L. M. (2022). Correcting a coarse-grid climate model in multiple climates by machine learning from global 25-km resolution simulations. *Earth and Space Science Open Archive*, 46. Retrieved from https://doi.org/10.1002/essoar.10511517.1 doi: 10.1002/essoar.10511517.1

Hodge, V., & Austin, J. (2004, October). A survey of outlier detection methodologies. *Artificial Intelligence Review*, 85–126. Retrieved from https://eprints.whiterose.ac.uk/767/ (Copyright © 2004 Kluwer Academic Publishers. This is an author produced version of a paper published in Artificial Intelligence Review. This paper has been peer-reviewed but does not include the final publisher proof-corrections or journal pagination.The original publication is available at www.springerlink.com.)

IPCC. (2021). Summary for policymakers [Book Section]. In V. Masson-Delmotte et al. (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* (p. 3 - 32). Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press. doi: 10.1017/9781009157896.001

Krasnopolsky, V. M., Fox-Rabinovitz, M. S., & Belochitski, A. A. (2010, July). Development of neural network convection parameterizations for numerical climate and weather prediction models using cloud resolving model simulations.

In *The 2010 international joint conference on neural networks (IJCNN)* (pp. 1–8). doi: 10.1109/IJCNN.2010.5596766

Krasnopolsky, V. M., Fox-Rabinovitz, M. S., Tolman, H. L., & Belochitski, A. A. (2008). *Neural network approach for robust and fast calculation of physical processes in numerical environmental models: Compound parameterization with a quality control of larger errors* (Vol. 21) (No. 2-3). doi: 10.1016/j.neunet.2007.12.019

Kwa, A., Clark, S. K., Henn, B., Brenowitz, N. D., McGibbon, J., Watt-Meyer, O., ... Bretherton, C. S. (2022, sep). Machine-learned climate model corrections from a global storm-resolving model: Performance across the annual cycle. Retrieved from `https://doi.org/10.1002/essoar.10512393.1` doi: 10.1002/essoar.10512393.1

Nairac, A., Townsend, N., Carr, R., King, S., Cowley, P., & Tarassenko, L. (1999, jan). A system for the analysis of jet engine vibration data. *Integr. Comput.-Aided Eng.*, *6*(1), 53–66.

Putman, W. M., & Lin, S.-J. (2007). Finite-volume transport on various cubed-sphere grids. *Journal of Computational Physics*, *227*(1), 55-78. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0021999107003105` doi: https://doi.org/10.1016/j.jcp.2007.07.022

Rasp, S., Pritchard, M. S., & Gentine, P. (2018). Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences*, *115*(39), 9684-9689. Retrieved from `https://www.pnas.org/doi/abs/10.1073/pnas.1810286115` doi: 10.1073/pnas.1810286115

Sanford, C. H., Kwa, A., Watt-Meyer, O., Clark, S., Brenowitz, N., McGibbon, J., & Bretherton, C. (2022). Improving the predictions of ml-corrected climate models with novelty detection. In *Neurips 2022 workshop on tackling climate change with machine learning.* Retrieved from `https://www.climatechange.ai/papers/neurips2022/10`

Schölkopf, B., Platt, J., Shawe-Taylor, J., Smola, A., & Williamson, R. (2001, 07). Estimating support of a high-dimensional distribution. *Neural Computation*, *13*, 1443-1471. doi: 10.1162/089976601750264965

Schölkopf, B., Smola, A., Williamson, R. C., & Bartlett, P. L. (2000). New support vector algorithms. *Neural Computation*, *12*, 1207-1245.

Sommer, C., Hoefler, R., Samwer, M., & Gerlich, D. W. (2017). A deep learning and novelty detection framework for rapid phenotyping in high-content screening. *Molecular Biology of the Cell*, *28*(23), 3428-3436. Retrieved from `https://doi.org/10.1091/mbc.e17-05-0333` (PMID: 28954863) doi: 10.1091/mbc.e17-05-0333

Song, H.-J., Roh, S., & Park, H. (2021, October). Compound parameterization to improve the accuracy of radiation emulator in a numerical weather prediction model. , *48*(20). doi: 10.1029/2021gl095043

Tan, Y., Niu, C., Tian, H., Hou, L., & Zhang, J. (2019). A one-class svm based approach for condition-based maintenance of a naval propulsion plant with limited labeled data. *Ocean Engineering*, *193*, 106592. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0029801819307188` doi: https://doi.org/10.1016/j.oceaneng.2019.106592

Van Aelst, S., & Rousseeuw, P. (2009, 07). Minimum volume ellipsoid. *Wiley Interdisciplinary Reviews: Computational Statistics*, *1*, 71 - 82. doi: 10.1002/wics.19

Wang, X., Han, Y., Xue, W., Yang, G., & Zhang, G. J. (2022). Stable climate simulations using a realistic general circulation model with neural network parameterizations for atmospheric moist physics and radiation processes. *Geoscientific Model Development*, *15*(9), 3923–3940. Retrieved from `https://gmd.copernicus.org/articles/15/3923/2022/` doi: 10.5194/gmd-15-3923-2022

Watt-Meyer, O., Brenowitz, N. D., Clark, S. K., Henn, B., Kwa, A., McGibbon, J., ... Bretherton, C. S. (2021). Correcting weather and climate models by machine learning nudged historical simulations. *Geophysical Research Letters*, *48*(15), e2021GL092555. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021GL092555` (e2021GL092555 2021GL092555) doi: https://doi.org/10.1029/2021GL092555

Woelfle, M. D., Yu, S., Bretherton, C. S., & Pritchard, M. S. (2018). Sensitivity of coupled tropical pacific model biases to convective parameterization in cesm1. *Journal of Advances in Modeling Earth Systems*, *10*(1), 126-144. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2017MS001176` doi: https://doi.org/10.1002/2017MS001176

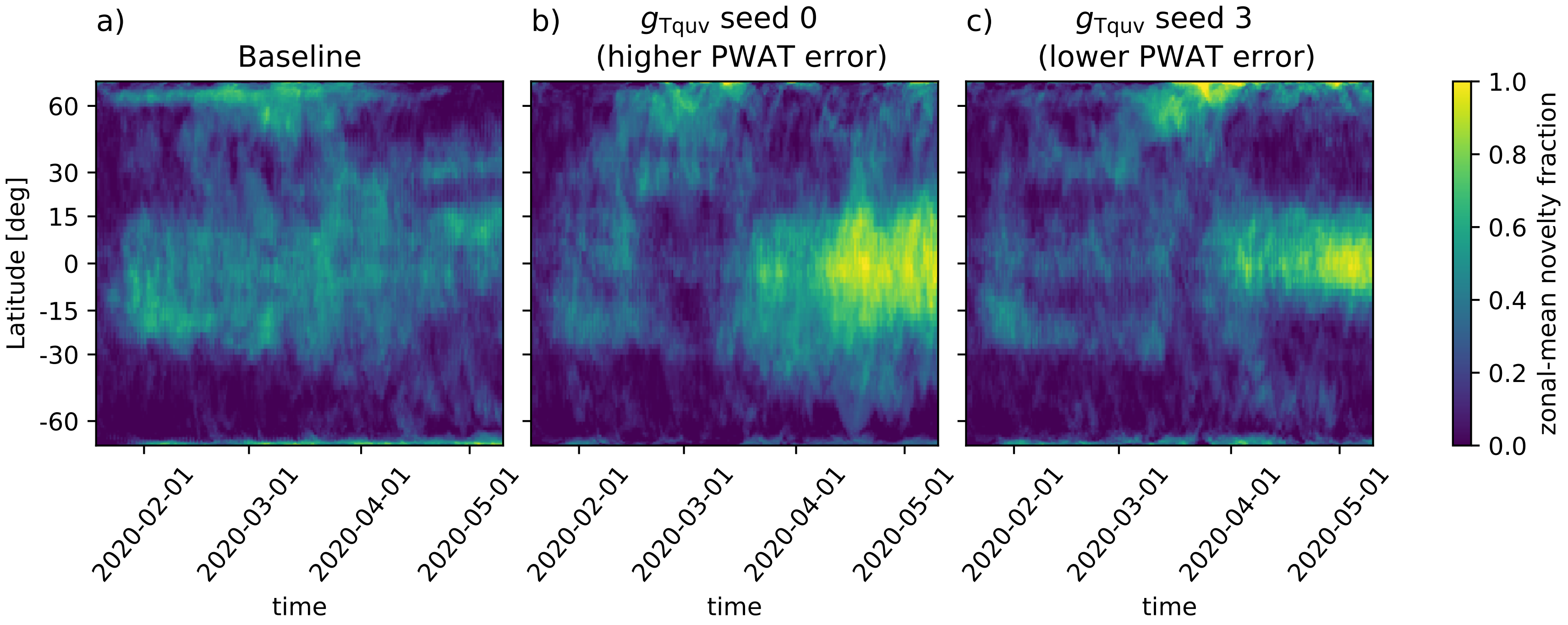
Yang, K., Kpotufe, S., & Feamster, N. (2021). An efficient one-class SVM for anomaly detection in the internet of things. *CoRR*, *abs/2104.11146*. Retrieved from `https://arxiv.org/abs/2104.11146`

Yuval, J., & O'Gorman, P. A. (2020). Stable machine-learning parameterization of subgrid processes for climate modeling at a range of resolutions. *Nature Communications*, *11*(1), 3295. Retrieved from `https://doi.org/10.1038/s41467-020-17142-3` doi: 10.1038/s41467-020-17142-3

Zhang, G. J., & Wang, H. (2006). Toward mitigating the double itcz problem in ncar ccsm3. *Geophysical Research Letters*, *33*(6). Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2005GL025229` doi: https://doi.org/10.1029/2005GL025229

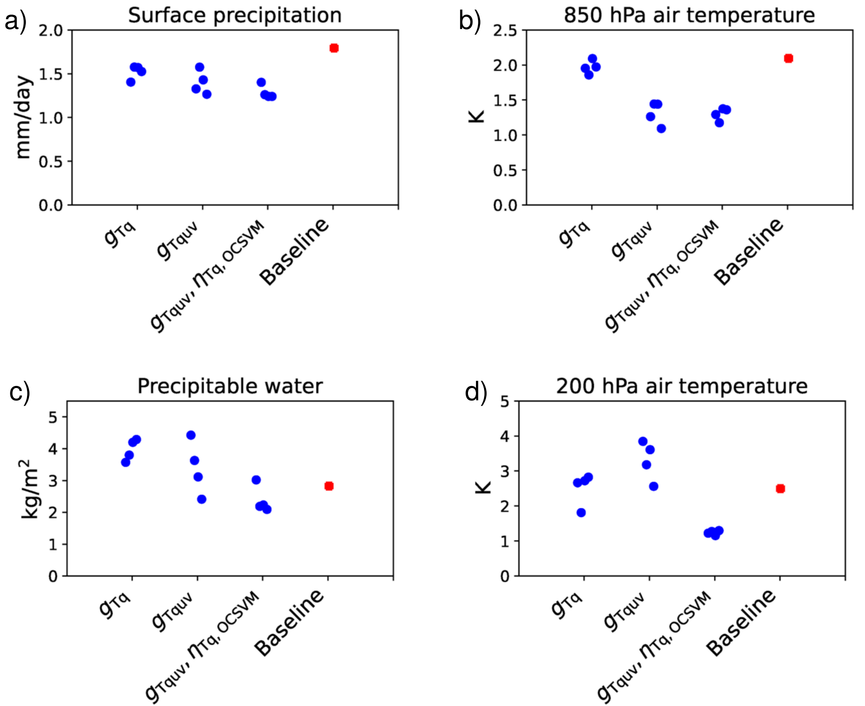
Zhou, L., Lin, S.-J., Chen, J.-H., Harris, L. M., Chen, X., & Rees, S. L. (2019). Toward convective-scale prediction within the next generation global prediction system. *Bulletin of the American Meteorological Society*, *100*(7), 1225 - 1243. Retrieved from `https://journals.ametsoc.org/view/journals/bams/100/7/bams-d-17-0246.1.xml` doi: 10.1175/BAMS-D-17-0246.1

Figure 1.



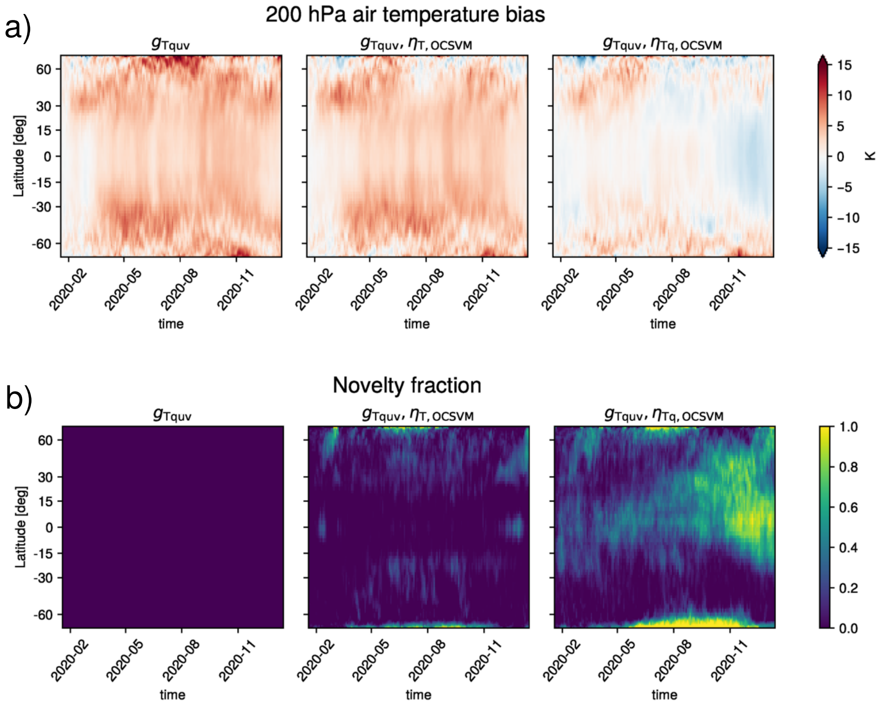
a)
Baseline
b)
g_{Tquv} seed 0
(higher PWAT error)
c)
g_{Tquv} seed 3
(lower PWAT error)
Latitude [deg]
60
30
15
0
-15
-30
-60
2020-02-01
2020-03-01
2020-04-01
2020-05-01
time
zonal-mean novelty fraction
0.0
0.2
0.4
0.6
0.8
1.0

Figure 2.



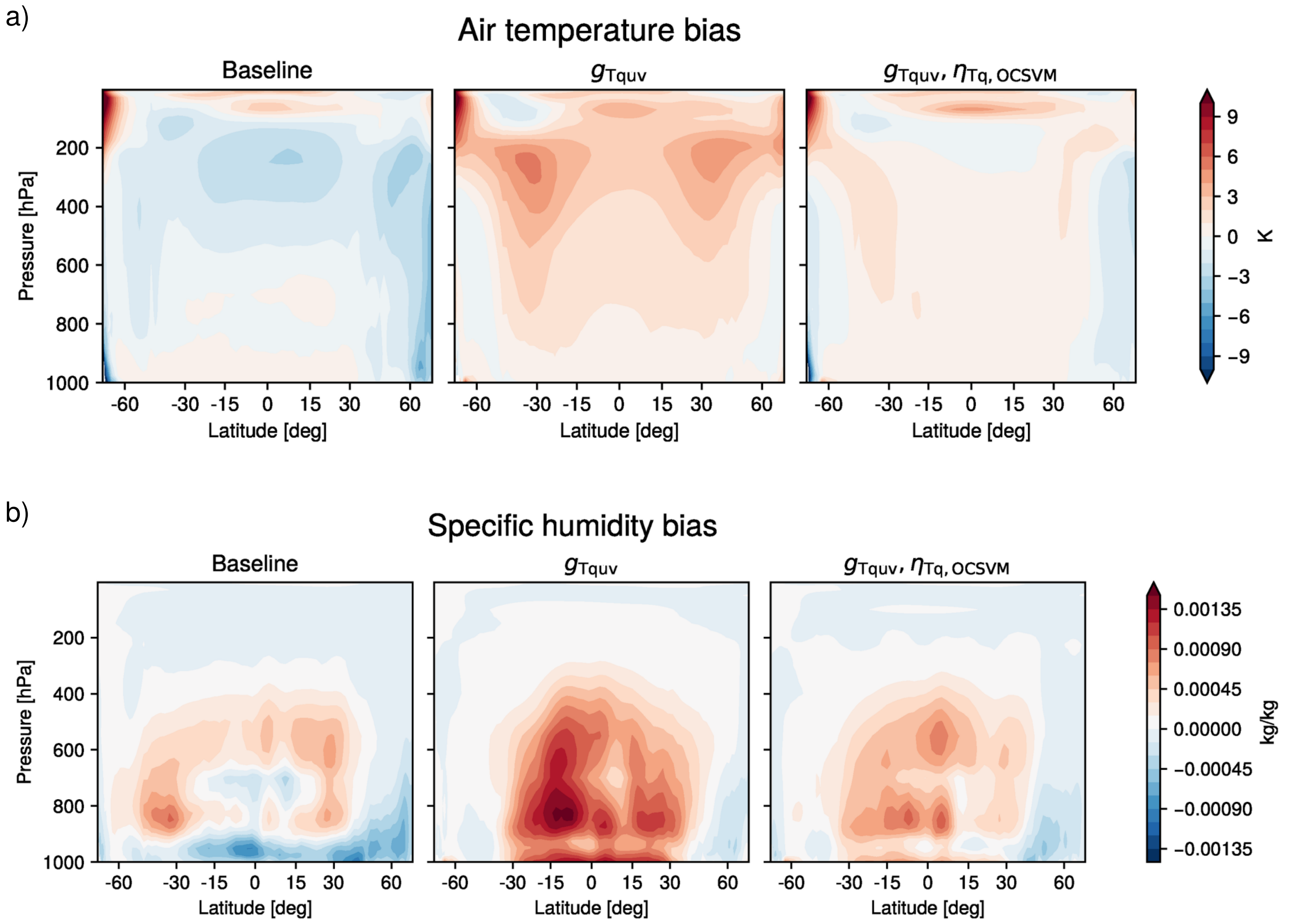
a)
Surface precipitation
mm/day
2.0
1.5
1.0
0.5
0.0
g_{Tq}
g_{Tquv}
$g_{Tquv}, \eta_{Tq,OCSVM}$
Baseline
b)
850 hPa air temperature
K
2.5
2.0
1.5
1.0
0.5
0.0
g_{Tq}
g_{Tquv}
$g_{Tquv}, \eta_{Tq,OCSVM}$
Baseline
c)
Precipitable water
kg/m^2
5
4
3
2
1
0
g_{Tq}
g_{Tquv}
$g_{Tquv}, \eta_{Tq,OCSVM}$
Baseline
d)
200 hPa air temperature
K
5
4
3
2
1
0
g_{Tq}
g_{Tquv}
$g_{Tquv}, \eta_{Tq,OCSVM}$
Baseline

Figure 3.



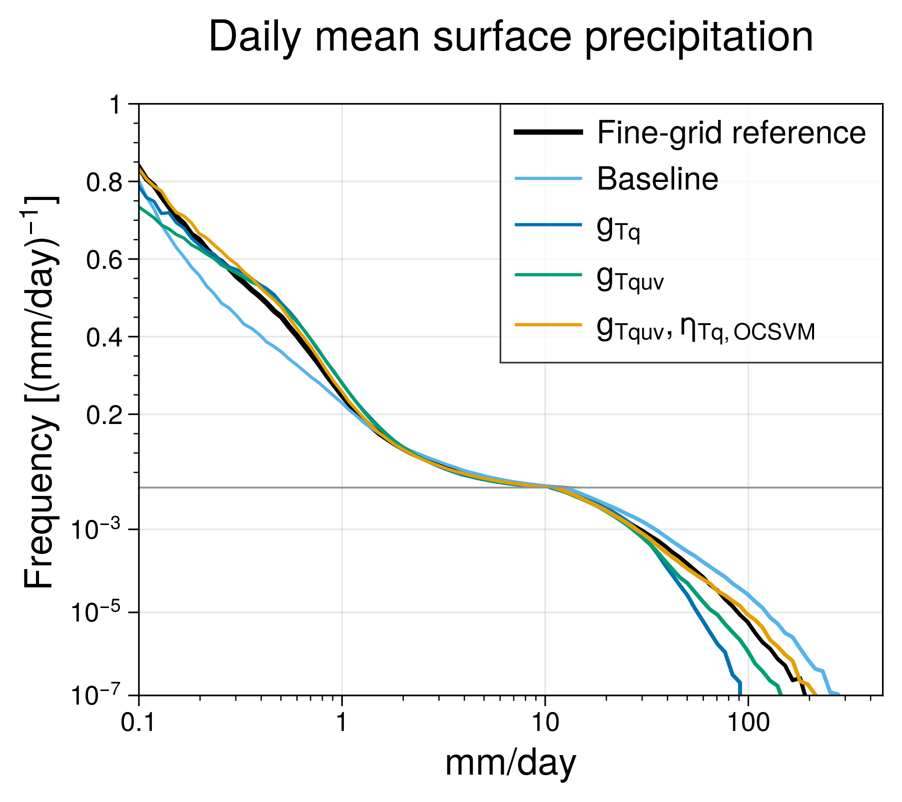
a)
200 hPa air temperature bias
g_{Tquv}
$g_{Tquv}, \eta_{T,OCSVM}$
$g_{Tquv}, \eta_{Tq,OCSVM}$
Latitude [deg]
time
K
b)
Novelty fraction
g_{Tquv}
$g_{Tquv}, \eta_{T,OCSVM}$
$g_{Tquv}, \eta_{Tq,OCSVM}$
Latitude [deg]
time

Figure 4.



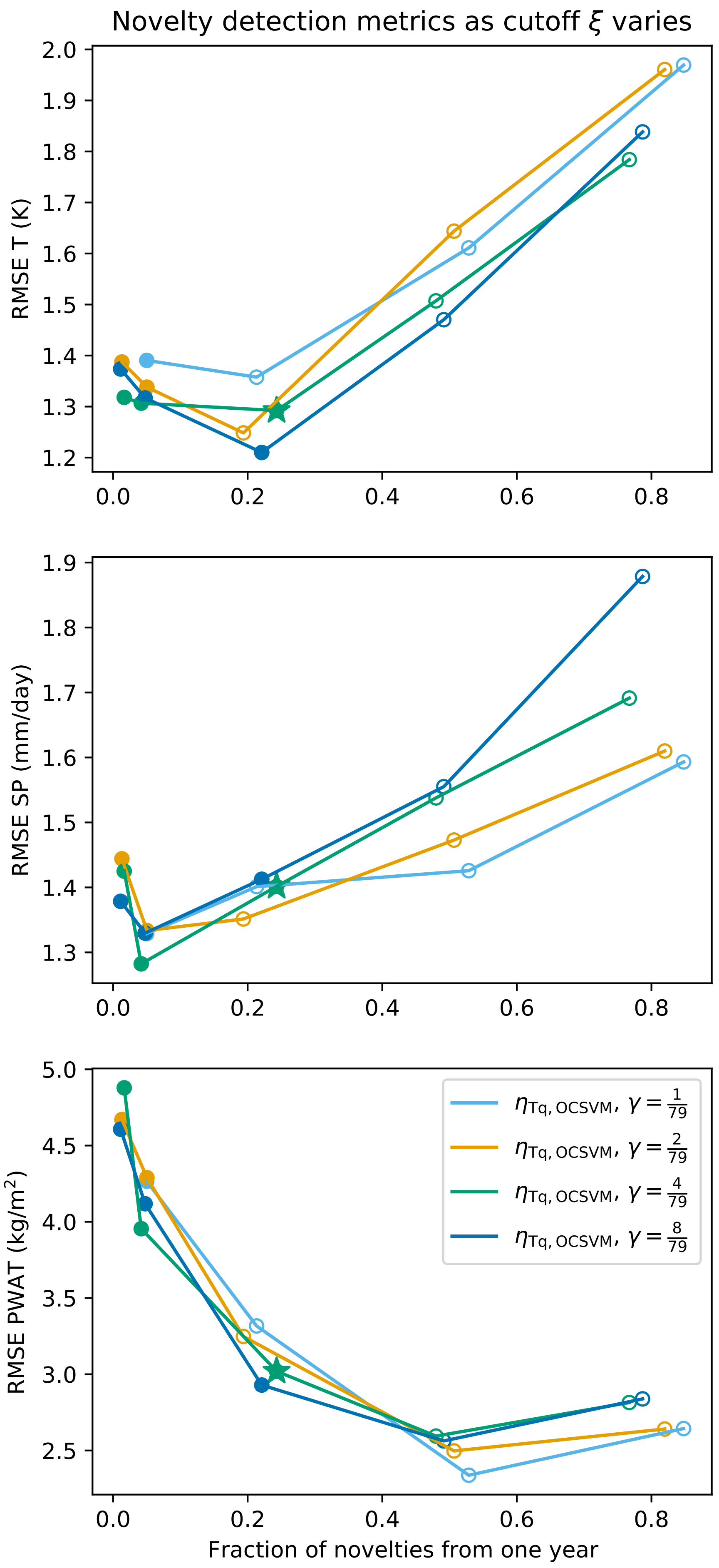
a)
Air temperature bias
Baseline
g_{Tquv}
$g_{\mathrm{Tquv}}, \eta_{\mathrm{Tq,\,OCSVM}}$
Pressure [hPa]
200
400
600
800
1000
-60
-30
-15
0
15
30
60
Latitude [deg]
K
9
6
3
0
−3
−6
−9
b)
Specific humidity bias
Baseline
g_{Tquv}
$g_{\mathrm{Tquv}}, \eta_{\mathrm{Tq,\,OCSVM}}$
Pressure [hPa]
Latitude [deg]
kg/kg
0.00135
0.00090
0.00045
0.00000
−0.00045
−0.00090
−0.00135

Figure 5.



Daily mean surface precipitation
Frequency [(mm/day)$^{-1}$]
mm/day
Fine-grid reference
Baseline
g_{Tq}
g_{Tquv}
$g_{Tquv}, \eta_{Tq,OCSVM}$
1
0.8
0.6
0.4
0.2
10^{-3}
10^{-5}
10^{-7}
0.1
1
10
100

Figure 6.



Novelty detection metrics as cutoff ξ varies
RMSE T (K)
2.0
1.9
1.8
1.7
1.6
1.5
1.4
1.3
1.2
0.0
0.2
0.4
0.6
0.8
RMSE SP (mm/day)
1.9
1.8
1.7
1.6
1.5
1.4
1.3
0.0
0.2
0.4
0.6
0.8
RMSE PWAT (kg/m^2)
5.0
4.5
4.0
3.5
3.0
2.5
0.0
0.2
0.4
0.6
0.8
Fraction of novelties from one year
$\eta_{\mathrm{Tq,OCSVM}}$, $\gamma = \frac{1}{79}$
$\eta_{\mathrm{Tq,OCSVM}}$, $\gamma = \frac{2}{79}$
$\eta_{\mathrm{Tq,OCSVM}}$, $\gamma = \frac{4}{79}$
$\eta_{\mathrm{Tq,OCSVM}}$, $\gamma = \frac{8}{79}$