Environmental Data Science (2022), 1: e11, 1–28
doi:10.1017/eds.2022.10

CAMBRIDGE UNIVERSITY PRESS

APPLICATION PAPER

A sensitivity analysis of a regression model of ocean temperature

Rachel Furner[1,2,*], Peter Haynes[1], Dave Munday[2], Brooks Paige[3], Daniel C. Jones[2] and Emily Shuckburgh[4]

[1]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, United Kingdom
[2]British Antarctic Survey, Cambridge, United Kingdom
[3]UCL Centre for Artificial Intelligence, Computer Science, University College London, London, United Kingdom
[4]Department of Computer Science and Technology, University of Cambridge
*Corresponding author. E-mail: raf59@cam.ac.uk

Received: 14 January 2022; **Revised:** 09 June 2022; **Accepted:** 21 July 2022

Keywords: Data science; interpretable ML; model sensitivity; oceanography; regression model

Abstract

There has been much recent interest in developing data-driven models for weather and climate predictions. However, there are open questions regarding their generalizability and robustness, highlighting a need to better understand how they make their predictions. In particular, it is important to understand whether data-driven models learn the underlying physics of the system against which they are trained, or simply identify statistical patterns without any clear link to the underlying physics. In this paper, we describe a sensitivity analysis of a regression-based model of ocean temperature, trained against simulations from a 3D ocean model setup in a very simple configuration. We show that the regressor heavily bases its forecasts on, and is dependent on, variables known to be key to the physics such as currents and density. By contrast, the regressor does not make heavy use of inputs such as location, which have limited direct physical impacts. The model requires nonlinear interactions between inputs in order to show any meaningful skill—in line with the highly nonlinear dynamics of the ocean. Further analysis interprets the ways certain variables are used by the regression model. We see that information about the vertical profile of the water column reduces errors in regions of convective activity, and information about the currents reduces errors in regions dominated by advective processes. Our results demonstrate that even a simple regression model is capable of learning much of the physics of the system being modeled. We expect that a similar sensitivity analysis could be usefully applied to more complex ocean configurations.

Impact Statement

Machine learning provides a promising tool for weather and climate forecasting. However, for data-driven forecast models to eventually be used in operational settings we need to not just be assured of their ability to perform well, but also to understand the ways in which these models are working, to build trust in these systems. We use a variety of model interpretation techniques to investigate how a simple regression model makes its predictions. We find that the model studied here, behaves in agreement with the known physics of the system. This works shows that data-driven models are capable of learning meaningful physics-based

This research article was awarded Open Data and Open Materials badges for transparent practices. See the Data Availability Statement for details.

patterns, rather than statistically valid but potentially spurious links, strengthening trust in the suitability of these systems for forecasting.

1. Introduction

1.1. Data-driven models for weather and climate

Applications of machine learning (ML) in weather and climate modeling, of both the ocean and the atmosphere, have seen a huge rise in recent years. Traditionally, weather and climate predictions rely on physics-based computational models of the Earth system, hereafter referred to as simulators or general circulation models (GCMs). Recently, a number of papers have focused on creating statistical/data-driven models for a variety of physical systems (Miyanawala and Jaimana, 2017; Pathak et al., 2018; Breen et al., 2020). These show the ability of statistics and ML to complement existing methods for predicting the evolution of a range of physical systems.

Lorenz models (Lorenz, 2006) are often used as a simple analogous system for weather and climate models as they have similar properties albeit in a considerably simplified way. Many data-driven models of the Lorenz equations have been developed and assessed (i.e., Dueben and Bauer, 2018; Chattopadhyay et al., 2019; Doan et al., 2019; Scher and Messori, 2019a). These results show that data-driven methods can capture the chaotic dynamics of the Lorenz system, and make skilled, short-term forecasts. A number of papers (Dueben and Bauer, 2018; Scher, 2018; Scher and Messori, 2019b; Weyn et al., 2019; Arcomano et al., 2020; Rasp and Thuerey, 2021) go further and apply statistical and ML methods to simple weather prediction applications, using a variety of model architectures, and training on both observational data and GCM output. Rasp et al. (2020) looks to standardize and formalize comparison of these methods. They propose a common dataset and test experiments, creating a common framework for assessing methods for predicting the short-term evolution of the atmosphere. The development of this field provides great promise for weather and climate predictions, with the demonstration of skillful forecasts, which could one day be used to provide efficient operational forecasts to complement existing physics-based GCMs.

1.2. Interpretable machine learning

Using data-driven methods in place of physics-based GCMs raises questions about how these models are making their predictions, and the reliability, trustworthiness, and generalizability of these models (McGovern et al., 2022). GCMs are based on known physics, meaning a single model can be used to reliably predict a variety of regimes. Data-driven models are instead dependent on the data used during training and the patterns learned by the model. The ability of a data-driven model to generalize, that is, to make skillful predictions for data which differs in some way to the data seen in the training set, depends on how the predictions are being made. If statistically robust patterns or links are found that hold well within the training data, but which ultimately have no physical basis, then we would not necessarily expect these models to perform well on data outside of the training set. For new examples, which bear little similarity to that seen in the training data and which are not close to any training examples in feature space, that is, extreme events not included in the training set, any nonphysical patterns that were learned are unlikely to hold and the model will not necessarily be expected to perform well. By contrast, if the performance over the training data is skillful because the model is learning meaningful physical links between the input and output variables, then we would expect the model to perform well for any data that exhibit these same physics, irrespective of the similarity to training samples. If data-driven methods for predicting weather and climate systems are able to learn the underlying dynamics of the system, rather than statistically valid but nonphysical patterns between inputs and outputs, we have increased confidence that these systems can be usable for a wide variety of applications.

https://doi.org/10.1017/eds.2022.10 Published online by Cambridge University Press

A number of techniques exist to understand the sensitivity of data-driven models, and to interpret how they are making their predictions, giving insight into their generalizability and reliability (Lipton, 2018; McGovern et al., 2019; Molnar et al., 2020). These techniques seek to help us understand not just whether a model is getting the right results, but if the models are getting the right results for (what we consider to be) the right reasons, that is, by learning meaningful physically consistent patterns. There are a number of model interpretation and visualization techniques, which focus on different elements of interpretability. Methods look at both identifying which features are important to a model (i.e., sequential search [Stracuzzi and Utgoff, 2004], impurity importance [Louppe et al., 2013; Breiman, 2001], permutation importance [Breiman, 2001]) and assessing how certain features are used by the model (i.e., partial dependence plots [Friedman, 2001] and saliency maps [Simonyan et al., 2013]). These methods seek to answer subtly different questions about how a model is working, and so it is common to use a variety of model interpretation techniques in parallel. Techniques that assess feature importance highlight which features are fundamental to the forecast, but not how these are being used. By contrast, methods that look at how features impact the forecast do not indicate the relative importance of these features for the predictions. As data-driven methods become more commonly used in weather and climate applications, so does an analysis into the interpretability of these models (McGovern et al., 2019; Mcgovern et al., 2020; Barnes and Barnes, 2021; Rasp and Thuerey, 2021).

1.3. Sensitivity study of a regression model of ocean temperature

The studies mentioned previously focus on atmospheric evolution, whereas, here, we focus on oceanic evolution. We develop a data-driven model to predict the change in ocean temperature over a day based on data from a GCM of the ocean, and then interpret this model through a variety of methods. The underlying physics explaining the dynamics of the Earth system is consistent across the atmosphere and the ocean. While there are many differences between atmospheric and ocean dynamics, for example, the temporal and spatial scales of interest, and compressibility of the fluid, these systems are driven by similar physics (Marshall et al., 2004). As such, the skills shown in using data-driven methods for predictions of the atmosphere (see references in Section 1.1) suggest that these same methods could provide skillful predictions for the evolution of the ocean.

The model developed here is highly simplified, both in terms of the idealized GCM configuration on which we train the model, and the data-driven methods used. However, the underlying configuration (Munday et al., 2013) captures key oceanic dynamics, enabling a suitable test bed to see if data-driven methods can capture the dynamical basis of these systems. Similarly, while we use a simple regression technique, this has sufficient skill to assess the ways in which the model works and to improve understanding of the potential of data-driven methods more generally.

We apply model interpretation techniques to our data-driven model to try to understand what the model is "learning" and how the predictions are being made, and compare this with our prior knowledge of the ocean dynamics. We analyze the sensitivity of the regressor to its input variables, firstly through a direct analysis of the coefficients of the resultant model to show which variables are heavily *used* in the forecasts, and secondly through withholding experiments to indicate which variables are *necessary* for producing skillful forecasts. Lastly, we further analyze some of the withholding experiments to infer *how* some of these key variables are contributing to the predictions.

Section 2 discusses methods: the GCM we use to create our training and validation dataset; the regressor we develop; and the sensitivity analysis we perform. Section 3 discusses the skill of the developed regressor. Section 4 explores the sensitivity of the regressor to its inputs. The results and their implications are discussed in Section 5.

https://doi.org/10.1017/eds.2022.10 Published online by Cambridge University Press

2. Methods

2.1. Simulator-generated dataset

2.1.1. Simulator configuration

Our training and validation data come from running the Massachusetts Institute of Technology general circulation model (MITgcm). This is a physically based model capable of simulating the ocean or the atmosphere due to isomorphisms in the governing equations (Marshall et al., 1997a,b). Specifically, we use a 2° sector configuration following Munday et al. (2013) to simulate ocean dynamics. This configuration features a single ocean basin, with limited topography, simplified coastlines, and constant idealized forcing. This has been used in a number of idealized simulations of Southern Ocean processes and their impacts on the global circulation (Munday et al., 2014). This configuration, while relatively simple, captures the fundamental dynamics of the ocean, including a realistic overturning circulation. The configuration is briefly described here, with key parameters given in Table 1. For further details, the reader is referred to Munday et al. (2013).

The domain runs from 60° S to 60° N, and is just over 20° wide in longitude. The domain is bounded by land along the northern (and southern) edge, and a strip of land runs along the eastern (and western) boundary from 60° N to 40° S (see Figure 1a). Below this, in the southernmost 20°, the simulator has a periodic boundary condition, allowing flow that exits to the east (west) to return to the domain at the western (eastern) boundary. The domain has flat-bottom bathymetry of 5,000 m over most of the domain, with a 2° region of 2,500-m depth at the southernmost 20° of the eastern edge (i.e., the spit of land forming the eastern boundary continues to the southern boundary as a 2,500-m high subsurface ridge).

The simulator has 42 (unevenly spaced) depth levels, following a Z-coordinate, with the surface layer being the thinnest at 10 m, and the bottom 10 levels being the maximum at 250 m. There are 11 cells in the longitudinal (x) direction, and 78 cells in the latitudinal (y) direction. The grid spacing is 2° in the latitudinal direction, with the longitudinal spacing scaled by the cosine of latitude to maintain approximately square grid boxes (this means that grid cells close to the poles are about a factor of 4 smaller in area than those near the equator, but all cells remain approximately square). The simulator has a 12-hr time step (two steps per day), with fields output daily. We focus on daily-mean outputs, rather than the instantaneous state.

Table 1. Key parameter information for MITgcm simulation.

Parameter	Value
Grid spacing (horizontal)	2°
Vertical levels	42 unevenly spaced vertical levels, with spacing ranging from 10 to 250 m
Harmonic viscosity (momentum)	$0.0075\mathrm{m}^4/\mathrm{s}$
Vertical viscosity (momentum)	$10^{-3}\mathrm{m}^2/\mathrm{s}$
GM coefficient	$1{,}000\mathrm{m}^4/\mathrm{s}$
Reference diapycnal diffusivity	$3\mathrm{e}^{-5}\mathrm{m}^2/\mathrm{s}$
Wind stress	$0.2\sin^2[\pi(\theta+60)/30]\mathrm{N/m}^2$ for $-60<\theta<-30$
Restoring timescale for salinity	30 days
Restoring salinity	$34+3/2(1+\cos(\pi\theta/240))$PSU
Restoring timescale for potential temperature	10 days
Restoring potential temperature	$30\sin[\pi(\theta+60)/120]$°C for $theta<0$ $5+25\sin[\pi(\theta+60)/120]$°C for $theta>0$

Abbreviations: GM, Gent–McWilliams; MITgcm, Massachusetts Institute of Technology general circulation model.

https://doi.org/10.1017/eds.2022.10 Published online by Cambridge University Press

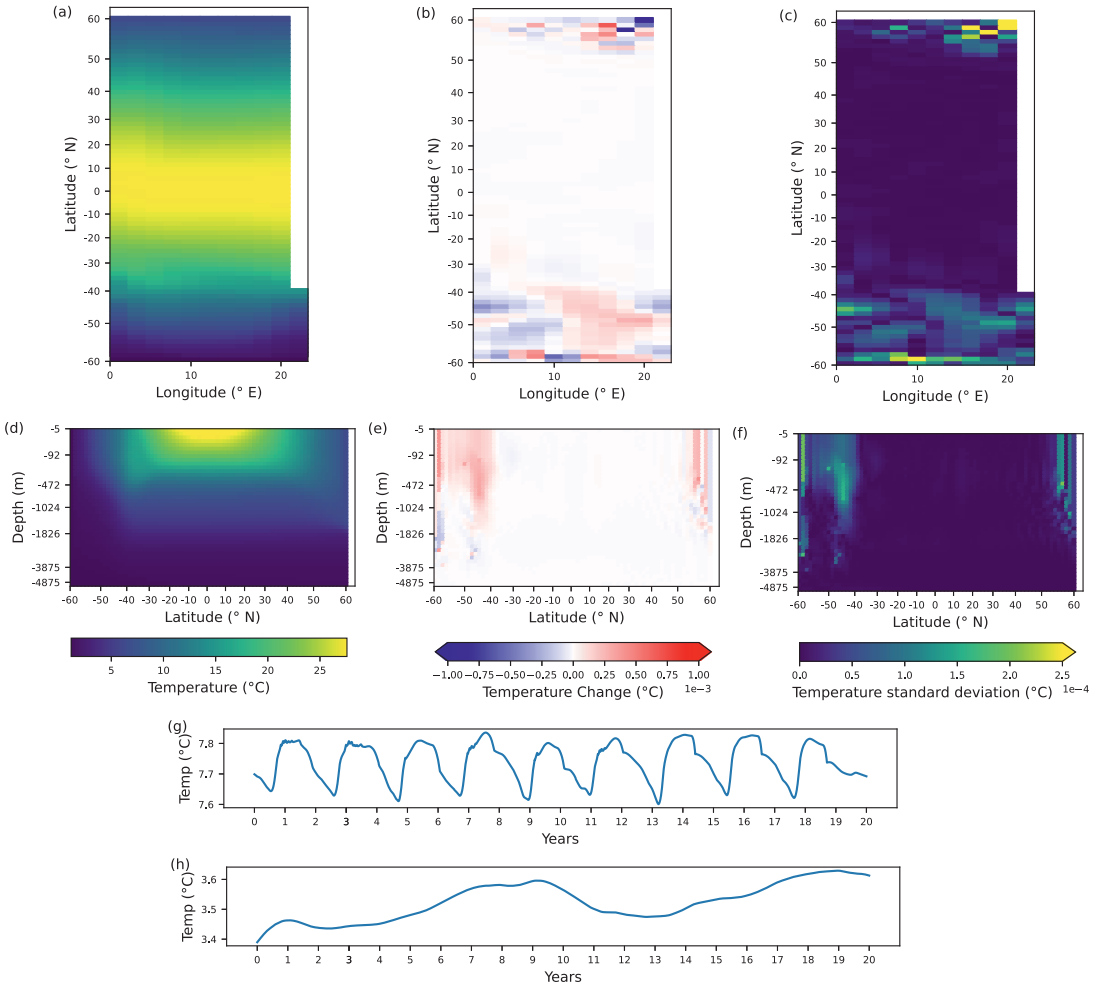

Figure 1. *Plot of simulator temperature (°C) at 25 m below the surface (a) and at 13° E (d), for one particular day. Change in temperature between over 1 day at 25 m below the surface (b) and at 13° E (e). Standard deviation in temperature at 25 m below the surface (c) and at 13° E (f). Time series at 57° N, 17° E, and −25 m (g), and at 55° S, 9° E, and −25 m (h). Note that the depth axis is scaled to give each GCM grid cell equal spacing. The simulator shows a realistic temperature distribution with warm surface water near the equator, and cooler water near the poles and in the deep ocean. Temperature changes are largest in the very north of the domain and throughout the southern region. Though changes per day are small, they accumulate over time to give cycles of around 0.2° in some regions of the domain.*

At 2° resolution, the simulator is not eddy-resolving, but uses the Gent–McWilliams (GM) parameterization (Gent and Mcwilliams, 1990) to represent the effects of ocean eddy transport. We ran the simulator with a strong surface restoring condition on Temperature and Salinity (see Table 1)—thus fixing the surface density. We applied simple jet-like wind forcing, constant in time, with a sinusoidal distribution (see Table 1) between 60° S and 30° S, with a peak wind stress value of 0.2 N/m^2 at 45° S.

2.1.2. Ocean dynamics

We are interested in predicting a change in temperature between two successive daily mean values. Figure 1a–f shows the daily mean temperature for a given day, along with the 1-day temperature change and the standard deviation in temperature across the 20-year simulation, for cross sections at 25-m depth

https://doi.org/10.1017/eds.2022.10 Published online by Cambridge University Press

and 13° longitude. Panels g and h show a time series of temperature for a point in the northeast of the domain, and a point in the Southern Ocean region.

From Figure 1, we can see that the simulator represents a realistic temperature distribution, with warm water at the surface near the equator and cooler water nearer the poles and at depth. The largest changes in temperature over a day are located in the south of the domain, and in a small region in the very north. These changes result predominantly from the local vertical activity associated with the meridional overturning circulation (MOC)—a wind- and density-driven circulation that is characterized by water sinking in the north, traveling southward at depth, and then upwelling in the south, where it splits in two, with some water returning northward near the surface, and some re-sinking to the south and returning north at depth (Talley, 2013; Rintoul, 2018). This circulation occurs on timescales of hundreds to thousands of years (this is the time taken for water parcels to complete one revolution); however, this large-scale circulation is the accumulation of the local vertical movements happening at very short timescales seen here. The largest daily temperature changes seen in the far north and south of this domain are the short timescale changes associated with this large-scale circulation. The MOC reflects the density profile, which itself arises from the surface forcing (the restoring term on temperature and salinity) and the wind forcing. Further details of the model dynamics, in particular assessment of the contribution of different processes to temperature change in the model, can be found in Appendix A. While we see from Figure 1b,e the daily changes in temperature are small, we see in Figure 1g,h that they accumulate to give far larger changes in temperature, up to 0.2°C per year, that is, on the order of 5×10^{-4}°C per day. As such, the changes that we see are significant in terms of the temperature variability that the model shows. Furthermore, predicting these small changes accurately is essential in enabling us to capture the larger trends happening over longer timescales, when using models that forecast iteratively over many time steps.

2.1.3. Training and validation datasets

Input and output data for training and validating the regressor come from a 70-year run of this simulator. The first 50 years of the run are discarded, as this includes a period where the model is dynamically adjusting to its initial conditions, which may be physically inconsistent. During this period, the evolution of the simulator is driven by this adjustment, rather than the more realistic ocean dynamics, which we are interested in; hence, we exclude these data. This leaves 20 years of data, which are used for training and validating the model. As the GCM sees a constant wind forcing and a consistent restoring of surface temperature and salinity, if left to run for long enough (thousands of years), the system would reach a quasi-steady state; however, the 20-year period used here is prior to the model reaching this quasi-steady state.

The data are highly autocorrelated, that is, fields are similar, particularly when considering fields that are temporally close. This strong autocorrelation, found in many weather and climate applications, impacts the ability of the algorithm. Therefore, as is common practice, we subsample in time to remove some of the codependent nature of the training data, optimizing the ability of the data-driven method. There are also computational constraints limiting the total size of our dataset. This leads us to choose a subsampling rate of 200 days, so every 200th field from the simulator is used in the dataset, and the rest discarded. This provides a balance between having large datasets (which in general benefit the algorithm), while also fitting within computational constraints, and limiting autocorrelation within the dataset. While this samples only around 40 temporal fields, the forcing is constant in our simulator and so we expect the dynamics to be reasonably consistent across time and therefore it is not necessary to sample across a large range of temporal fields.

To clarify, the subsampling is the time between sample fields used to train and validate the model. This is in no way connected to the time of the prediction step, which in this work is 1 day. While acknowledging that there is little variation in the dynamics over time, we still expect that the temperature change between Day 0 and Day 1 is similar to the temperature change between Day 1 and Day 2, but less similar to the temperature change between Day 200 and Day 201. In order for any data-driven method to learn well, we need to provide a set of training samples that is large enough, and which are as different as possible, and

https://doi.org/10.1017/eds.2022.10 Published online by Cambridge University Press

adequately sample the variation in 1-day temperature change. Here, we ensure this by taking our first sample from fields at $t = 0$ and $t = 1$, our second sample from fields at $t = 200$ and $t = 201$, and so forth.

This dataset is then split into training, validation, and test data with a split of 70–20–10. The data are systematically split temporally, so the first 70% of samples are used as training data and so forth, meaning that each dataset contains data from different temporal sections of the run, maximizing independence across the datasets. For every 200th pair of days, we take all grid points from the model interior. We exclude points next to land and points at the surface and seabed, as the algorithm developed here is not suitable for forecasting these points—the regressor requires input from surrounding points, and so is only suitable for predicting the interior of the domain. We do not subsample in space, as the domain is reasonably small and the dynamics varies considerably across it, meaning that subsampling in space can lead to some dynamic regimes being entirely missing from the dataset. This gives us approximately 650,000 training samples, 200,000 validation samples, and 100,000 test samples.

2.2. Regression model of ocean temperature

We develop a regression model to predict the daily mean change in ocean temperature for any single grid cell, based on variables at surrounding locations at the current time step. The regressor is defined such that it outputs temperature change at a single grid cell rather than predicting for the entire domain, but the cell being predicted can be any of the cells in the domain interior—the regressor is not limited to predicting for a specific location.

Equation (1) shows the formulation of the regressor.

$$\hat{y} = \sum_{i=1}^{N_f} \beta_i x_i + \sum_{\substack{i,j=1 \\ i<j}}^{N_f} \gamma_{i,j} x_i x_j. \tag{1}$$

Here, $\hat{y}$ is the output from the regressor—an estimate of the change in daily mean temperature over a day for the grid cell being predicted. This is calculated as the mean temperature at the next day $(t + 1)$ minus the mean temperature at the present day (t). N_f is the number of input features used in the model. β_i and $\gamma_{i,j}$ are the weights of the regressor, which are learnt during the training phase. x_i and x_j are the input features being used to make the predictions.

Input variables are temperature, salinity, U (East–West) and V (North–South) current components, density, U, V, and W (vertical) components of the GM bolus velocities, sea surface height (SSH), latitude, longitude, and depth. The GM bolus velocities are a parameterization of the velocities resulting from ocean eddies and are used in the GM scheme to calculate the advective effects of eddy action on tracers. For 3D variables (temperature, salinity, current components, density, and GM bolus velocity components), input features are taken from a $3 \times 3 \times 3$ stencil of grid cells, where the center cell is the point for which we are predicting, giving 27 input features for each variable. For SSH, which is a 2D variable, the values over a 2D (3×3) stencil of surrounding locations are included, giving a further nine features. Lastly, the location information (latitude, longitude, and depth) at only the point we are predicting is included, giving the final three input features. All temporally changing variables are taken at the present day (t). In total, this gives $N_f = 228$ features, represented by the first term in equation (1).

We also include second-order pairwise polynomial terms, in order to capture a limited amount of nonlinear behavior through interaction between terms. This means that as well as the above inputs, we have multiplicative pairs of features, represented by the second term in equation (1). Note that we include second-order interactions between different features, but not squared terms, as we are interested in representing the interaction between different features through this term. This gives 26,016 input terms in total.

The model design means that all physical ocean variables at surrounding points are included in the prediction, as these are likely to impact the change in temperature at the central point. Geographic inhomogeneity in the dynamics is accounted for through inclusion of the location information.

https://doi.org/10.1017/eds.2022.10 Published online by Cambridge University Press

Furthermore, the combination of this geographic inhomogeneity with physical ocean variables is included to a limited extent through some of the multiplicative terms in equation (1) (those terms that are a combination of latitude, longitude, or depth with a physical variable input). Lastly, the nonlinear interaction between physical ocean variables is also included to a limited extent through the remainder of the multiplicative terms. All input variables are normalized prior to fitting the model by subtracting their mean and dividing by their standard deviation.

2.2.1. Limitations of the model

It should be noted that the model is a simple regressor, to allow for easy analysis of sensitivity. This, however, limits how accurately the model can fit the data, and how well it can represent the underlying system. In particular, we know the ocean to be highly nonlinear, but allow only second-order polynomial terms in the regressor, restricting the level to which it can capture the true dynamics.

The regressor here takes input data from only immediately surrounding grid cells, meaning that it has no information about what is happening in the wider domain. This potentially prevents the regressor from making predictions far ahead, when the wider ocean state has more influence, but for the short time steps being forecast here (1 day), this local information is expected to be sufficient. Indeed, here, we are making predictions at time steps only double that used in the GCM—where the change at each cell is based predominantly on the state of only immediately surrounding cells.

Lastly, we note that many existing papers looking at data-driven forecast systems focus on developing methods that can be applied iteratively to provide an alternative forecast system able to predict an arbitrary number of time steps ahead. However, the model described here would not, in its current form, be usable to produce an iterative forecast in this same way. Our work is motivated by these examples of data-driven models that are used iteratively to produce a forecast, but our interest is not in deriving a data-driven analog of the MITgcm simulation, which might one day be used in place of the original simulator, but simply in assessing the sensitivity of a data-driven model to different variables. Focusing our sensitivity analysis on single time-step predictions means that we remain focused on the sensitivity of the model directly, rather than any potential artifacts of the forecast associated with the iteration. The inability of the model to iterate is therefore not an issue for the focus of this work.

There are two reasons why our existing setup is unable to iteratively forecast. First, the regressor requires a wider set of inputs than the outputs it produces, and so iterative forecasting would require some means of generating variables other than temperature to provide the full set of inputs to the regressor at all time steps, that is, we would require a number of regression models, forecasting all variables. As our focus on model sensitivity is best addressed through focusing on a single variable with a single model, we do not attempt that here. Second, this model is unable to forecast near the boundary as it requires a full set of neighboring input points. We chose to focus our work on an ocean application; however, this introduces the additional challenge of dealing with a land–sea interface. To the best of our knowledge, this has not yet been approached from a data-driven perspective. As the focus of this work is on assessing the sensitivity of the model, we chose not to attempt solutions to this problem here, but instead to work with a model suitable for the ocean interior only.

We believe that focusing on a single variable and using an easily interpretable data-driven model best allows us to assess the dependencies and sensitivities of an example data-driven model. Furthermore, while our model is not capable of iterating, the analysis carried out and the conclusions around the sensitivity and trustworthiness of our model are still relevant to the wider discussion of sensitivity of data-driven models.

2.2.2. Training the regressor

The model is trained by minimizing least-squared errors with ridge regularization (Hoerl and Kennard, 1970). Training a standard least-squared model amounts to finding values of the coefficients (β_i and $\gamma_{i,j}$), which minimize the squared difference between the regression model predictions and the actual outputs taken from the GCM over the training dataset. In any application of a regression model, it is expected that

https://doi.org/10.1017/eds.2022.10 Published online by Cambridge University Press

the model will be used on data other than that used in the training of the model. To ensure that the model performs well on unseen data, we want to ensure that that model learns the general pattern of the data, rather than specifically fitting to every point in the training data. This is particularly important where datasets are known to contain noise, as here fitting the data exactly would mean "learning" the noisy representation of the system that the data portrays, rather than learning the underlying system itself. Regularization techniques are applied to avoid the problem of overfitting (of matching the training data exactly) and work to limit the level at which the model can fit the data, ensuring that the model can generalize well—that is, it still performs well on new unseen data that share the same underlying dynamics. Ridge regression is one such regularization method, which works by minimizing the size of the coefficients as well as the model errors. When using ridge regression, an additional term is added to the loss function, so the training now focuses on minimizing a combination of the square errors and the sum of the magnitude of the β_i and $\gamma_{i,j}$ values, with α acting as a tuning parameter determining the balance between these two terms.

We use a very small value of α=0.001. This was found through cross-validation with the values of α ranging from 0.001 to 30. With larger values, the regressor performed poorly, particularly when predicting larger temperature changes. Given that the dataset comes from simulator output, we know that, in this case, noise or measurement error is not an issue, so the need for regularization is limited. Similarly, while we have a large number of weights in our equation, the size of our training set is very large compared with this, which already acts to limit overfitting. Because of this, we find that only very small values are necessary.

2.3. Sensitivity studies

We wish to investigate the sensitivity of the regressor to its inputs in order to understand the ways in which the regressor is making its predictions. We do this in three ways. First, we directly assess the coefficients (weights) used in the resulting regressor. This indicates which features are being most heavily *used* in the predictions. Second, we run a series of withholding experiments, and this indicates which inputs are most *necessary* for accurate forecasts. Lastly, for the inputs that the withholding experiments identified as being most critical to forecasts, we assess the impact these have on errors, giving insight into *how* these inputs effect the forecasts.

We assess the coefficients simply through plotting a heat map of coefficients (Figure 4 and Section 4.1). Inputs that are highly weighted by the regressor (those with large coefficients) are important to the prediction, whereas those with low weights can be considered as less important for the predictions.

Alongside this, we run a series of withholding experiments (Table 1 and Section 4.2). For each of the variables described in Section 2.2, with the exception of temperature, we train a new regressor leaving out that one variable group, for example, we train a new regressor with all the existing inputs except for salinity at all surrounding points and any multiplicative terms including salinity. This corresponds to running the first pass of a Backward Sequential Search interpretability analysis. We also run two further withholding experiments. In the first, we assess the importance of providing information in the vertical neighborhood of points. Instead of the 3D stencil originally used, we take a 2D neighborhood of points (3×3) in only the horizontal direction, thus giving nine inputs for each of temperature, salinity, and so forth. Lastly, we also run without multiplicative terms, that is, the model consists of only the first term in equation (1), giving a purely linear equation, enabling us to assess the impact of nonlinearity on predictions. The new regressors are trained in exactly the same way, using the same training and validation samples—the only difference being the number of input features used. Comparing results from these withholding experiments to the control run show the importance of the withheld variable—if error increases significantly, then the variable is necessary for accurate predictions. However, if the impact on error is small, the regressor is able to make predictions of similar accuracy with or without that variable, indicating that it is not needed for good predictions.

While these two methods (coefficient analysis and withholding experiments) help to indicate the feature importance in the model, it should be noted that they highlight different aspects of the importance

https://doi.org/10.1017/eds.2022.10 Published online by Cambridge University Press

of the input features. Looking at the coefficients of the trained regressor helps to identify which inputs are being most heavily *used* for the predictions from that particular regressor. By contrast, the withholding experiments indicate which variables are *necessary* to get predictions with the level of accuracy shown in the control. There may, for example, be scenarios where certain variables are heavily weighted and flagged as important through the coefficient analysis, but when these same variables are withheld, the regressor re-weights other variables during the training step and maintains a similar level of accuracy due to correlations and the strong codependency of ocean dynamics on multiple variables. Coefficient analysis helps us to understand how a particular instance of a regressor is working, whereas the withholding experiments help us to understand the impact and importance of each variable in creating skillful regression models more generally.

Lastly, we analyze the resultant models from the three worst-performing withholding experiments. We look at scatter plots of truth against prediction and spatial plots of averaged absolute error to see how these models perform. We compare the average error plots to average errors in the control run (a run with all inputs) to see where errors are increased. We then compare this with the dominant processes driving temperature change in those regions (Figures A1 and A2) and our expectations based on prior knowledge of ocean dynamics to assess if the regressors respond in the ways we expect.

3. Performance of the Regressor

First, we discuss the performance of the control model—the regressor which is trained using the full set of previously discussed inputs. The predictions from the regression model closely resemble the true change in daily mean temperature in both the training and validation datasets (Figure 2) although there is a tendency to underpredict the magnitude of temperature changes.

The model captures the central part of the distribution well. While the majority of the temperature change is dominated by small near-zero changes, capturing these is key to producing a good forecast system. Although the complete development of a data-driven forecast system is not the focus of this work, we are motivated by the potential for data-driven methods to replicate traditional forecast systems. As such, the ability of the model developed here to capture the full range of dynamic behavior, beginning with the most common dynamics, is key.

To a lesser extent, the regressor also captures the tails of the distribution, where temperature changes are larger, although the underprediction is more significant here. However, it is noteworthy that the model still shows some skill for these points, given that the model used is very simple and there are a relatively limited number of training samples in the tails—of the over nearly 650,000 training samples, just over 500 of those samples have temperature changes in excess of ± 0.001°C. Despite the relatively rare nature of these larger temperature changes, we feel that capturing these alongside the smaller changes is important in building a robust model. The underlying dynamics of the system, which we hope the regression model is able to learn, drives the full range of temperature changes seen. As such, if we build a regressor which is unable to capture the extreme levels of change, this would indicate that the model is not fully learning the physical dynamics as was intended. Capturing these extremes is also critical to obtaining a model which could (with further development) lead to a feasible alternative forecast system. Given the simplicity of the regressor used here, it is promising that it captures the extremes to the limited extent shown. However, the results also identify the need for more sophisticated methods that can better capture both the dominant dynamics and the extreme cases.

Table 2 reports root-mean-square (RMS) errors for this run (top row) in comparison with a persistence forecast (bottom row). A persistence forecasting is a forecast of no change—in this case, to forecast zero temperature difference. It is important to consider RMS errors in relation to a benchmark forecast, to distinguish between the difficulty of the problem being studied and the skill of the model being used. Persistence forecasts are commonly used as a benchmark in forecasting and provide a statistically good predictor for this problem due to the limited temperature change across most of the simulator domain. However, we can see that the regressor performs significantly better than persistence. As expected, we can

https://doi.org/10.1017/eds.2022.10 Published online by Cambridge University Press

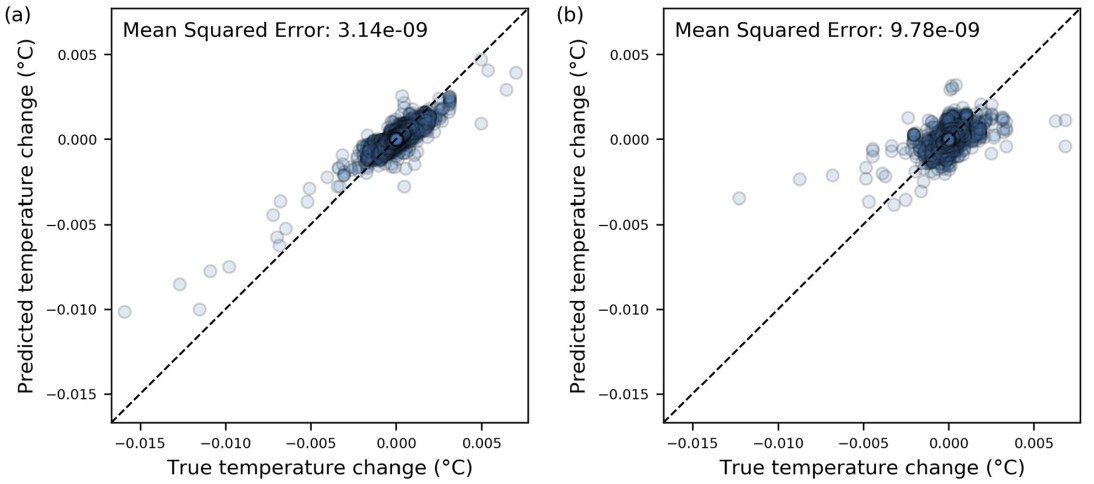

Figure 2. *Scatter plot of predictions against truth for both training (a) and validation (b) datasets for the control regressor. Over the training set, the regressor does a good job of predicting for both the dominant near-zero behavior, and the very rare temperature changes of more than ±0.002°. Over the validation dataset, the regressor drops in accuracy, with a tendency to underpredict, particularly for large changes, but still shows some skill.*

Table 2. Table showing RMS errors and RMS errors normalised by the control for a series of withholding experiments. Results are also included from a persistence model (bottom row) for comparison. The two withholding experiments which make the largest difference to each error metric are shown in italic. These are ordered in terms of RMS error over the training dataset, with variables which are most necessary for predictive skill appearing nearest the bottom. It is critical to include polynomial interactions. Information on the vertical structure, and on the currents is also necessary for good predictive skill.

	RMS error (°C)		RMS error normalized by control	
Experiment	Training	Validation	Training	Validation
Control (full inputs set)	5.61e-05	9.89e-05	—	—
Withholding longitude	5.65e-05	9.92e-05	1.01	1.00
Withholding depth	5.66e-05	9.91e-05	1.01	1.00
Withholding latitude	5.66e-05	9.94e-05	1.01	1.01
Withholding salinity	5.82e-05	1.01e-04	1.04	1.02
Withholding density	5.82e-05	1.02e-04	1.04	1.03
Withholding SSH	5.89e-05	1.01e-04	1.05	1.02
Withholding bolus velocities	7.32e-05	9.65e-05	1.30	0.98
Withholding currents	8.16e-05	*1.07e-04*	1.45	*1.08*
Using a 2D (3 × 3) input stencil	*8.52e-05*	1.06e-04	*1.52*	1.07
Without polynomial interactions	*1.02e-04*	*1.14e-04*	*0.12*	*0.11*
Persistence model (for comparison)	1.02e-04	1.15e-04	1.82	1.16

Abbreviation: SSH, sea surface height.

https://doi.org/10.1017/eds.2022.10 Published online by Cambridge University Press

see from Table 2 and Figure 2 the regressor performs less well over the validation dataset; however, it consistently outperforms the persistence forecast.

Anomaly correlation coefficients on the predicted field (i.e., over the predicted temperature, T, rather than the predicted temperature increment, δT) were also calculated, giving values of 0.9999987 and 0.9999916 over the training and validation datasets, respectively. Anomaly correlation coefficients values are frequently reported in papers that develop data-driven models (i.e., Scher and Messori, 2019b; Rasp et al., 2020), hence their inclusion here. However, it should be noted that it is not trivial to compare these statistics across differing applications, as the results are heavily influenced by the difficulty of the problem being addressed, rather than purely indicating model skill. For this work, we do not feel correlation coefficients to be a useful metric and focus instead on RMS errors.

3.1. Spatial patterns of errors

We calculate temporally averaged absolute errors to give us an indication of how the regression model performs spatially. These averages were created by taking the MITgcm state at 500 different times from the 20-year dataset and using these fields as inputs to the regressor to forecast a single time step ahead. The set of forecasts created from these 500 input states is compared to the truth from the GCM run, and the absolute errors between the truth and the predictions are then temporally averaged. To emphasize, this is an average of 500 single time-step predictions, and not an average from an iterative run.

The set of input states spans the full 20-year MITgcm dataset, but with subsampling to take every 14th day (as opposed to every 200th day as was used in creating the training and validation sets). This results in a far larger set of input states than present in the training and validation data. The results here are therefore not specific to either the training or validation set, but instead show performance over a larger dataset which shares occasional samples with both.

These averaged errors are shown in Figure 3. Note that the regressor is only applied away from boundary and land points (in its current form, it cannot deal with spatial locations that are not surrounded on all sides by ocean points); hence, points close to land are not included in these plots.

Figure 3 shows the largest errors are located in the north of the domain and in the Southern Ocean. These are regions where the temperature change itself is largest (compared with Figure 1, which shows snapshots of daily temperature change) as would be expected. In particular, the large errors throughout the Southern Ocean section of the domain persist through depth, although the largest errors are associated with points above 1,000 m, or at the very southern extent of the domain.

Comparing Figure 3b with Figures A1 and A2, we see that the errors in the north of the domain are co-located with regions of high vertical advective temperature fluxes, and regions of high convective fluxes. These results imply the regression model struggles to fully capture the vertical processes, and the associated heat flux, in the north of the domain. The high errors in the Southern Ocean are again co-located with regions of high vertical diffusive fluxes, this time both explicit and implicit, and vertical advection. However, the pattern is less clear here, as the location of these errors is also a region of high meridional diffusive fluxes and high zonal advective fluxes. Throughout the ocean interior where temperature changes and the fluxes associated with these are small, errors are also small as would be expected.

The results are promising given the limitations of this model. Although we allow second-order polynomial interactions, we are still working with a very simple regression model, and the order of complexity is nowhere near that considered to be present in the simulator, or the physical ocean. To truly capture the dynamics of the ocean, far higher levels of interaction and complexity would be required. That a simple regressor achieves this level of skill is promising when considering the potential for applications of more complex data-driven methods, such as the neural networks described in Dueben and Bauer (2018), Scher (2018), Weyn et al. (2019), Arcomano et al. (2020), and so forth.

https://doi.org/10.1017/eds.2022.10 Published online by Cambridge University Press

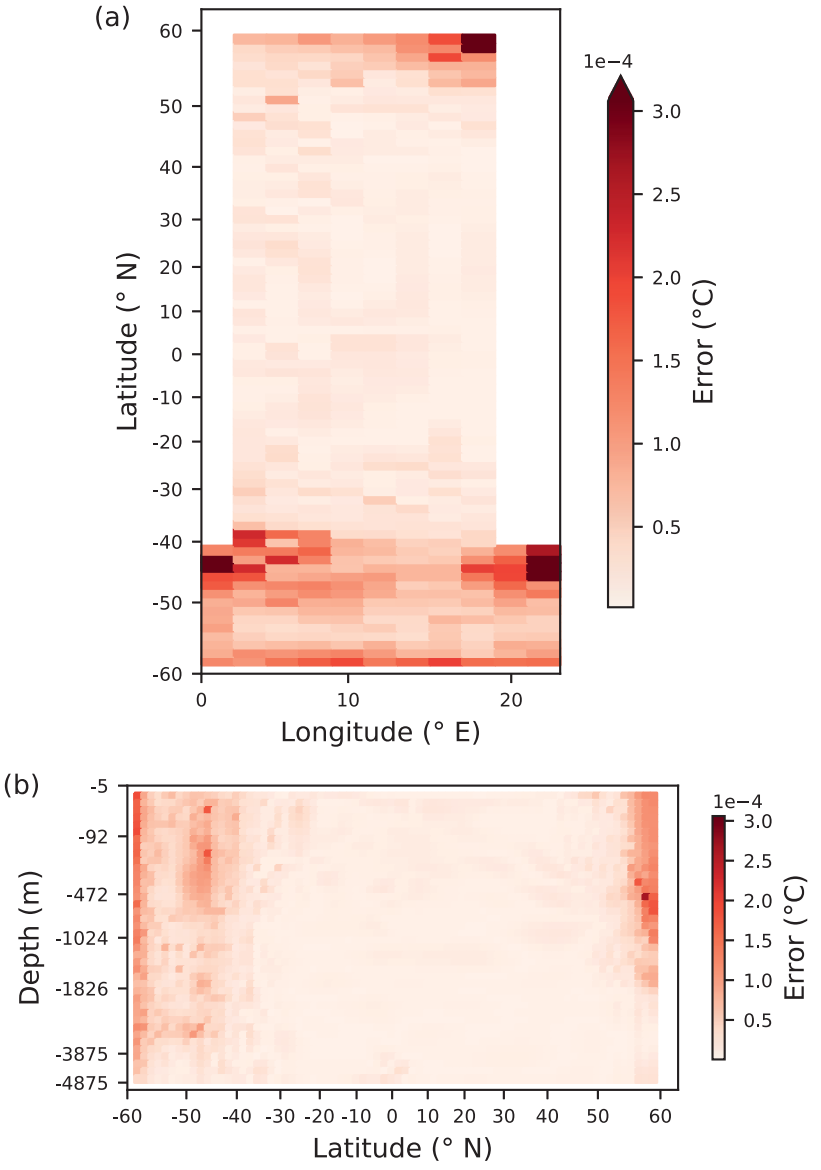

Figure 3. *Mean Abs Error of predictions (°C) at −25 m depth (a) and 13° E (b).The errors are largest in the very north of the domain, and in the southern region, in locations where the temperature change itself is largest. Comparing with Figures A1 and A2, we see that errors are largest in the areas of increased vertical fluxes and locations with high meridional diffusion, and high zonal advection.*

4. Sensitivity of the Regressor

4.1. Coefficient analysis

First, we assess the sensitivity of the trained regressor by direct coefficient analysis. Figure 4 plots the magnitude of the coefficients in equation (1). Figure 4a shows coefficients averaged over all input locations for each variable type (i.e., for most variables, there are 27 inputs, corresponding to the 27 neighboring cells; we average over these to give a single value for each variable (temperature, salinity, etc.) and for each polynomial combination of variables). Figure 4b shows the coefficients related to polynomial interactions of temperature with temperature—these are the raw coefficients, without any

https://doi.org/10.1017/eds.2022.10 Published online by Cambridge University Press

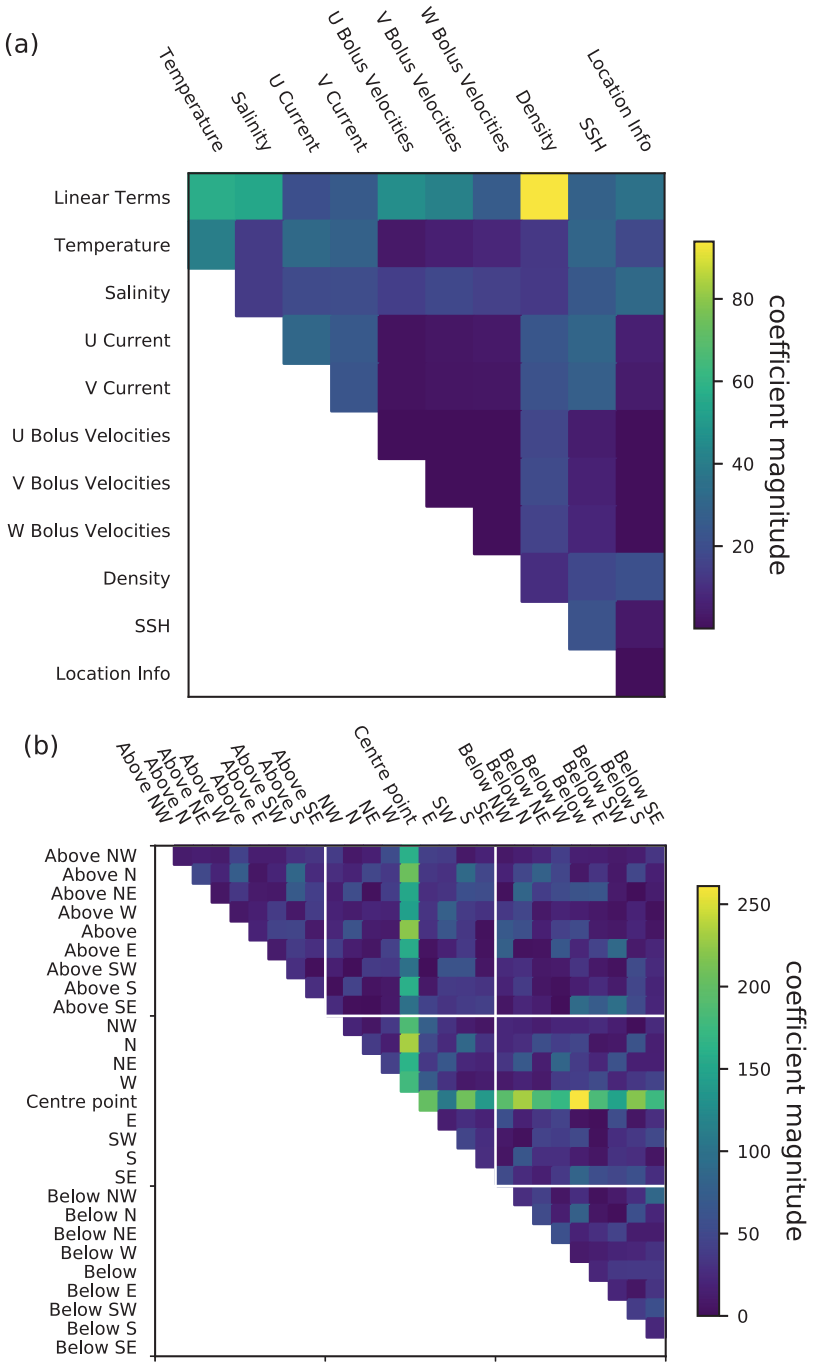

Figure 4. *(a) Coefficients of the control regressor. Coefficients averaged over all input locations for each variable type, and each set of nonlinear combinations of variables. (b) Coefficients for polynomial terms representing temperature–temperature interactions across all pairs of input locations. We see that density is very heavily weighted, and therefore providing a large part of the predictive skill of this model, this is in line with our physical understanding that density changes are driving convective temperature change. The interactions between the temperature at the point we are predicting and the temperature at surrounding points are also very highly weighted. This is in line with our physical knowledge of advection and diffusion driving temperature change.*

https://doi.org/10.1017/eds.2022.10 Published online by Cambridge University Press

averaging applied. Supplementary Fig. B1 shows the full set of coefficients without any averaging applied. High-weighted inputs (those with a large magnitude coefficient) are variables which are heavily used in the predictions and therefore considered important for the predictive skill, whereas inputs with low magnitude coefficients can be considered less important. Again, we emphasize that the coefficients highlight which features are being predominantly *used* in this model, but this is not necessarily what is *needed* to create a skillful model—for that, we need to look at the withholding experiments.

From Figure 4a, we see that density (as a linear term, not in combination with other variables) is by far the most highly weighted variable in this model. The regressor is using density information as a very large component of its predictions. This is promising, as from our physical understanding of the system, we know that density is key to ocean dynamics. Unstable density profiles contribute to the large temperature changes seen in the south and very north of the domain, and for geostrophic currents, the flow follows the density stratification.

More generally, we see that the location information is low weighted, particularly when interacting with other variables. This indicates that the regressor is not basing its predictions predominantly on the location of points, but on the physical variables themselves.

From Figure 4b, we see that the multiplicative interaction between temperatures at different input locations is very highly weighted for certain combinations of locations. Specifically, it is the interaction between the temperature at the grid point we are predicting for and the temperature at all surrounding points, which gives the bright banding. This fits well with our physical expectation of the system—as diffusive and advective fluxes of temperature are dominated by local gradients in temperature.

4.2. Withholding experiments

RMS errors from a series of withholding experiments are shown in Table 1, along with results from the control and a persistence forecast. Withholding experiments quantify the relative necessity of each input variable. The larger the increase in error between the control and a withholding experiment, the more necessary the withheld feature is for making accurate predictions. All withholding experiments perform at least as well as the persistence model (which is used as a benchmark in weather and climate models) over the training and validation datasets, indicating that even with incomplete input sets the regression models developed here show significant skill.

4.2.1. Withholding location information

The inputs that have the smallest impact on training error are those giving location information about the grid point being predicted (the longitude, latitude, and depth of the grid cell). These variables have no direct influence on the dynamical processes driving temperature changes in the simulator (note that while latitude has physical relevance in ocean dynamics due to it being directly linked with the Coriolis effect, this does not directly drive temperature change—its impacts appear through changes in velocities, which are provided to the regressor already). That the regressor performs well even when the model has no location information indicates that well-performing regressors are not heavily dependent on learning patterns that are non-physically based on location, but may instead be learning patterns based on the underlying dynamics.

4.2.2. Withholding physical variables

The physical ocean variables have higher impacts on errors than the location variables—indicating that the regressor requires knowledge of the physical system in order to make its predictions. Of these, withholding salinity, density, or SSH information has minimal impact. Again, these variables have limited direct influence on temperature—their effects are felt through the resulting changes in currents caused by interactions of these variables. In a model able to capture more complexity, or looking at forecasting over longer time periods, these variables may become more relevant; however, when looking at evolution of

https://doi.org/10.1017/eds.2022.10 Published online by Cambridge University Press

just temperature over a single day, they are of little direct importance, both physically and when developing skillful regression models.

While density was a heavily weighted coefficient, when withholding density, the impact is small, especially when compared with the impact of currents. This highlights the usefulness of interpreting models through a variety of techniques, each of which gives insight into different aspects of the way the model is working. The density of seawater depends on its temperature and salinity, and so is tightly coupled to both of these. While the control model used density strongly in making its predictions, when density is withheld, the model has the ability to adjust by using these tightly coupled variables more heavily, enabling it to still provide accurate predictions. This tight coupling and interdependency of density with other variables likely explains the small impact seen in the withholding experiments. The combination of information from the two methods used to analyze feature importance indicates that density information is very highly used by the model when available, but that its usefulness can be easily compensated by other variables if it is not provided to the model, that is, it is sufficient but not necessary for model skill.

The experiment withholding information about the currents performs the worst of all the experiments concerning physical variables. That currents are one of the most important inputs required for regressor performance implies that some understanding of advection in the regression model is critical for accurate results, in line with our knowledge of the physical system being modeled. Errors from this experiment are analyzed in more detail in Section 4.3.

4.2.3. Withholding vertical structure and multiplicative terms

The withholding experiments which have the highest impact on training error are those which train on only a 2D stencil, or include only linear terms. Again, these experiments are analyzed in further detail in Section 4.3.

Using a 2D stencil means the regressor has no information about the ocean vertically above and below the location being predicted, and cannot use the vertical structure of the ocean in its prediction. We know this information to be important in the dynamics of the simulator, particularly in the south of the domain and the very north where vertical processes driven by the MOC affect temperature, and so it is reassuring that withholding it has such a large impact on error.

By restricting the regressor to purely linear terms (withholding polynomial interactions), we see the largest increase in error over the training set. That this purely linear version of the regressor performs poorly is also expected given our physical understanding of the problem being modeled. The ocean is known to be a complex, highly nonlinear system, and we would not expect a purely linear regressor to be able to accurately replicate the detail and variability of these complex interactions.

4.2.4. Summary of withholding experiments

These withholding experiments emphasize that in order to provide even a basic level of skill in forecasting temperature change in the ocean, a regression model needs information on currents and vertical structure, as well as enough complexity to capture some of the nonlinearity of the system. The feature importance displayed here by the regressor is consistent with the importance of these inputs in the dynamic system we are modeling, implying that the model is dependent on the variables we would expect. Therefore, we are confident that the regressor is, to some extent, learning physical constraints rather than purely statistical patterns that might lack causality.

4.3. Further analysis of withholding experiments

We further investigate the results of the three worst-performing models from the withholding experiments; withholding information on the currents, providing only 2D inputs, and a purely linear model. We look closely at the model predictions and errors, and compare these with the control run to infer how the variables are impacting predictions.

https://doi.org/10.1017/eds.2022.10 Published online by Cambridge University Press

4.3.1. Impact of multiplicative terms

Figure 5 shows the performance of the purely linear model, that is, the model trained without any multiplicative terms. We see that, without multiplicative terms, the model can capture the mean behavior of the system (zero change in temperature) but is unable to capture any of the variability. This mean behavior alone does not provide useful forecasts, as can be seen from the statistics for this experiment. Comparing Figure 5 with Figure 2, we see the importance of the nonlinear terms in predicting temperature change, especially for samples where temperature change is nonzero. Nonlinearity is shown to be critical to modeling the variability of temperature change.

4.3.2. Impact of vertical structure

To assess how information about the vertical structure of the ocean impacts predictions, we look at spatially averaged errors from the model trained with only a 2D neighborhood of inputs, along with the difference in error between this and the control run (Figure 6). Figure 6a is created in the same way as Figure 3b, with the absolute error from predictions across the grid at 500 different times averaged to give a spatial pattern of errors. Figure 6b shows the difference between Figures 3b and 6a, with areas shaded in red indicating where the error has increased as a consequence of withholding information about the vertical structure, and blue indicating areas where the predictions are improved. By comparing Figure 6b with Figures A1 and A2, we can see which processes are dominant in the regions of increased error, and make inferences about the ways in which the additional inputs are being used in predictions.

Interestingly, this regressor shows some regions (the deep water in the south of the domain) where the errors are notably improved in a regressor using only 2D information. In this work, we have developed a regressor which learns one equation to be applied across all grid boxes in the domain. We optimize for best performance averaged over all relevant grid cells, but this does not enforce the best possible performance over each individual grid point/region, and so some of the resultant models will favor certain types of dynamics more than others. Given this, it is not unexpected that the new equations discovered for the withholding experiments (which again optimize for best performance averaged over the entire domain interior) may outperform the control in some locations, despite being poorer overall. Here, we see that the control model is able to perform well across the domain, and optimizes for good performance overall (see Figure 3b), rather than the much more varied performance seen in the withholding experiments

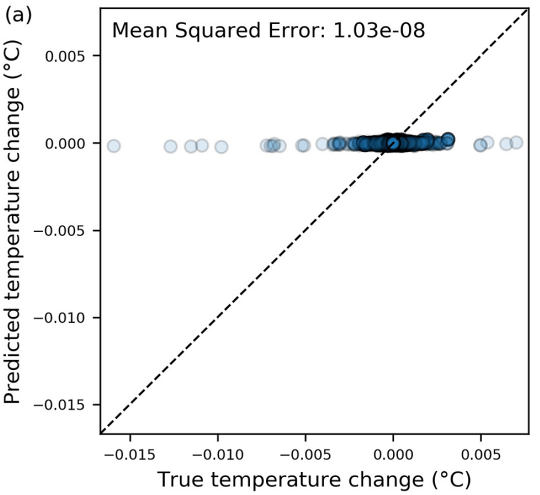

Figure 5. *Scatter plot of predictions against truth over the training dataset for the regressor trained with no polynomial interaction terms. A purely linear regressor (trained without nonlinear interactions) is unable to capture the behavior of the system. This is expected as we know the underlying system to be highly nonlinear.*

https://doi.org/10.1017/eds.2022.10 Published online by Cambridge University Press

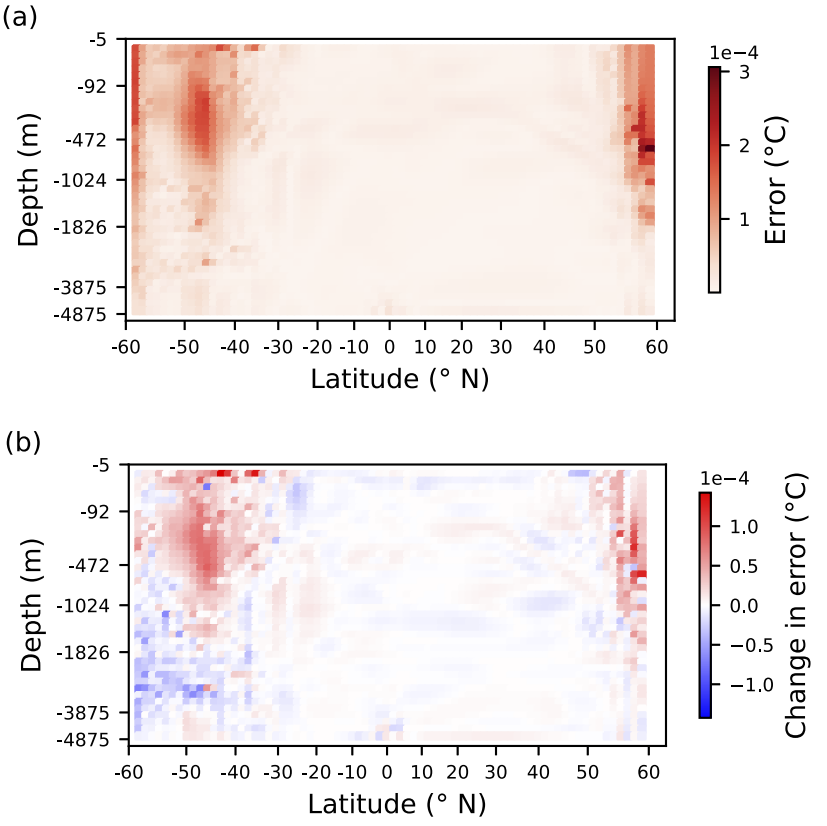

Figure 6. *(a) Cross section at 13° E of Mean Abs Error for the regressor trained using a 2D stencil. (b) Difference between this and the control run (Figure 3b). When withholding information about the vertical structure, errors in the regressors prediction are increased in a region north of 50° and south of −30°. Comparing this with Figures A1 and A2, we can see how the areas of increased errors correspond to particular processes.*

(Figure 6b). It seems that as the model which withholds vertical information is not capable of performing well in many regions of the domain, a solution is found which highly optimizes performance in other regions to minimize error overall.

This highlights the limitations of our method, and the potential need for more complex data-driven models that can better adjust to the wide variety of dynamics shown across the domain. It would be possible to produce a plethora of simple regression models, each of which is optimized for different locations within the domain, and combine these to produce a domainwide prediction. However, this would be a far more computationally demanding challenge, and would bring with it large risks of overfitting. With this sort of design, each regional model, seeing only a subset of dynamics, may be less likely to "learn" the underlying dynamics of the ocean, and instead learn statistically accurate but dynamically less-valid local patterns. However, other more sophisticated modeling methods could be explored to find a single model which has the complexity to better capture the detailed nonlinear dynamics in the ocean.

More interestingly, we see that using a 2D stencil rather than a 3D stencil increases errors in the very north of the domain, and in a region south of −40°. The area of increased error in the north coincides most closely with a region of high convective fluxes. We note that it also corresponds to a lesser extent with a region of high vertical advection; however, the shape and the location near the surface seem to far better correspond with the region of high convection. Convective activity is driven by dense water overlying less

https://doi.org/10.1017/eds.2022.10 Published online by Cambridge University Press

dense water leading to vertical mixing. For the regressor to "learn" the change in temperature associated with this, it would require information about the vertical density profile. That errors are increased in this region when information about the vertical structure is withheld implies that the model is dependent on the vertical structure in the ways we would expect.

The increased errors seen in the upper waters of the Southern Ocean are more complicated. They are roughly co-located with regions of high zonal advection and high meridional diffusivity. This is unexpected, given that these are horizontal processes and should not depend on the regressor having knowledge of what is happening above and below the point being predicted. However, we can see from Figures A1 and A2 that the Southern Ocean is a region of very complex dynamics (considerably more so than other regions in this GCM configuration), with many different processes occurring. Within this complex dynamical region, there are clear signals of high vertical diffusive fluxes and convection, which would be more in line with our physical expectations, although these appear far broader than the specific regions of increased error which we see. It may be that the increase in errors in this region is driven by the regressors reduced ability to capture the vertical diffusion and convection, as would be in line with our physical expectations. However, these results more strongly indicate that the regressor is learning spurious links between the inputs provided for a vertical neighborhood of points, and zonal advection and meridional diffusion. It should be emphasized that the complex dynamics of the Southern Ocean may test the limitations of such a simple regressor, causing the model to revert to less-physically relevant patterns in this area. In particular, in this region, currents flow along non-horizontal isoneutral surfaces, meaning that there is inherent interaction between the processes considered here. It may well be the case that such a simple model is not able to capture this interaction, and a similar assessment performed on a more complex data-driven model would be of interest here.

It is important to emphasize that this analysis only infers plausible explanations, but it is not able to definitively attribute the increased errors to any specific process. We see here that there are very plausible explanations for the errors seen in the north of the domain, which are in line with what we expect from a regressor which has learned the underlying dynamics of the ocean. By contrast, while there are physically consistent explanations available for the increased errors in the south of the domain, there are stronger indications of less physically consistent behavior. This implies that in the complex Southern Ocean region, the regressor struggles to fully capture the dynamics of the region, particularly with regard to the way it uses information on the vertical structure of the ocean.

4.3.3. Impact of currents

We analyze the impact of the currents on the regressor by again looking at the locations where errors are most changed between this experiment and the control run, and comparing these to the dominant processes in those areas. Figure 7 shows the spatially averaged errors from this regressor along with the difference between these and the errors from the control model. Again, we see a small number of points where errors are reduced with the simplified model. This is for the same reasons as described in Section 4.3.2.

The horizontal (U and V) components of the currents directly drive horizontal advection of temperature. They are also indirectly related to horizontal diffusion, as this is increased in regions of high currents and steep gradients. As such, we would expect that suppressing information about the horizontal currents would cause increases in error in regions where horizontal advection and horizontal diffusion is high. Comparing Figure 7b to Figures A1 and A2, we do indeed see a region of increased error south of −40°, which coincides with the regions of high zonal advection and high meridional diffusivity. However, again, we note that this region of increased error is one where many processes are present, and the increased errors seen also coincide, to a lesser extent, with regions of high vertical processes (advection, diffusion, and convection), which is less in line with our physical understanding. Here, errors appear more closely matched to the horizontal processes, and so a reasonable interpretation is that the model here is behaving as expected, although again we emphasize that it is not possible, based on the evidence here, to definitively attribute the increased errors to any specific process, only to make plausible inferences.

https://doi.org/10.1017/eds.2022.10 Published online by Cambridge University Press

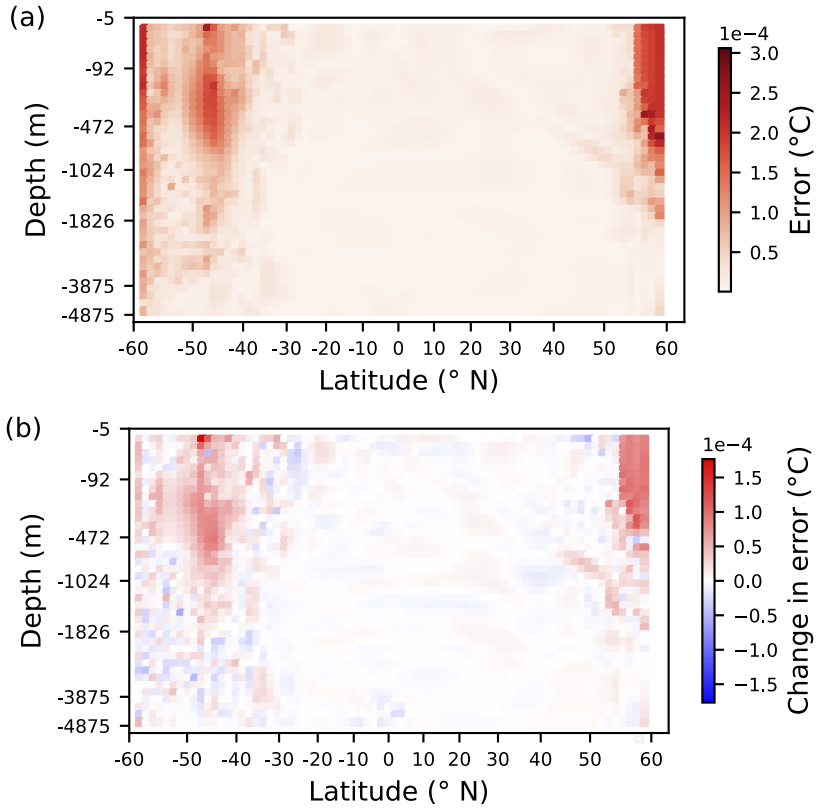

Figure 7. *(a) Cross section at 13° E of Mean Abs Error for the regressor trained with information on the currents withheld. (b) Difference between this and the control run (Figure 3b). When withholding currents, errors in the regression model are increased north of 55°, and in a broader region south of −35°. Comparing this with Figures A1 and A2, we can see how the areas of increased errors correspond to particular processes.*

The largest increase in errors are in the very north of the domain—an area where the temperature flux is dominated by vertical processes, both vertical advection (driven by vertical currents) and convective activity (i.e., due to instabilities in the water column). The increased errors in this northern region seen in Figure 7b seem to most closely correspond with the region of large vertical advection seen in Figure A1c. While it may at first be counter-intuitive that errors are increased in a region dominated by *vertical* advection when *horizontal* currents are withheld, this is in fact in line with our understanding of the dynamics of the system. Vertical advection is indirectly linked to the horizontal currents, as vertical currents are predominantly a consequence of convergence or divergence of the horizontal flow (particularly as the vertical motion of the water resulting from unstable density profiles manifests in the convective fluxes). The results here imply that as the regressor is not directly given information on the vertical currents, it may be learning the link between the horizontal and vertical currents, and the resultant vertical advection. Without information on the horizontal currents, the regressor struggles to capture this vertical advection resulting in increased errors in this northern region, in line with our understanding of the physical processes being modeled. It is noteworthy that the increase in errors here are larger than those in Figure 6. However, if our hypothesis is correct, that the errors here are associated with vertical advection and that the errors in Figure 6 are associated with convection, then the different contributions to heat flux of these two processes (see the scales in Figures A1 and A2) explain the smaller change in errors seen here.

https://doi.org/10.1017/eds.2022.10 Published online by Cambridge University Press

5. Conclusions

There is growing interest in the potential for ML to provide data-driven weather and climate forecasts, as an alternative to traditional process-based GCMs. A number of recent examples show these models to perform well in predicting the short-term evolution of the atmosphere (Dueben and Bauer, 2018; Scher, 2018; Scher and Messori, 2019b; Weyn et al., 2019; Arcomano et al., 2020; Rasp and Thuerey, 2021). However, alongside more standard performance metrics (RMS error, correlation coefficients, etc.), an understanding of the generalizability and trustworthiness is key to acceptance and use of any ML model. There are many studies of the interpretability of data-driven models in the geosciences more broadly (McGovern et al., 2019; Mcgovern et al., 2020; Barnes and Barnes, 2021). Specifically, focusing on data-driven forecast models for weather and climate, Rasp and Thuerey (2021) present a data-driven forecast model and interpret this using saliency maps. They show that in some cases, the model behaves in unexpected ways, highlighting the need for a thorough assessment of how these models work before they might be more widely accepted and used by the geoscience community. We continue to address this question of generalizability and trustworthiness of data-driven forecasts by assessing the sensitivity of a simple ocean model.

We have developed a simple regression model to predict the evolution of ocean temperature. Despite being a simple statistical tool, the developed model is able to predict change in daily mean temperature from an ocean simulator with notable skill when appropriate inputs are provided. That such a simple data-driven method can make skillful predictions gives promise to the growing set of data-driven approaches for weather and climate modeling. One concern around these methods is the lack of physical basis that might limit the ability for these models to perform well "out of sample" (i.e., over datasets outside of the training region). For the regression model developed here, we have shown that the sensitivity of the model outputs to the model inputs is generally in line with our physical understanding of the system.

Specifically, we analyze the coefficients of our regression model and find that the predictions for a grid cell are based heavily on the density at the surrounding points, and the interaction between the temperatures at the grid cell and its neighboring points. The importance of temperature interaction with surrounding points is representative of advective and diffusive processes that take place across the domain. The importance of density is in line with the simulator representing, to some extent, density-driven currents that are responsible for much of the changes in temperature in this GCM configuration. While later withholding experiments show that density is not *necessary* for skillful predictions, this is most likely due to the dependency of density on temperature and salinity, and the regressors ability to use these variables in place of directly using density when density is not available as an input. Again, this behavior makes sense when considering the physics of the ocean.

We conduct a number of withholding experiments. These show that withholding information about the location of the grid cell being forecast has very little impact on accuracy. In contrast, withholding information on the physical ocean variables has a larger impact. Of these, the velocities have the biggest impact, in line with our knowledge of advection being a key process in the transfer of heat. We see that inclusion of nonlinear interactions between inputs, and information about the vertical structure (rather than solely the horizontal structure), are both needed for skillful predictions. Again, this is compatible with our knowledge of the physical system. The ocean is highly nonlinear, and it would be expected that a nonlinear model is needed to capture its behavior. Similarly, the ocean dynamics are inherently three-dimensional, and so it is expected that inputs from a 3D neighborhood are necessary for predictive skill.

Further analysis of the three worst-performing withholding experiments give insight into how these inputs impact predictions. We see that including some level of nonlinearity is critical to capturing the complex nature of the system. Looking spatially at the errors from experiments that withhold currents, and withhold information about the vertical structure, we see that errors are generally increased in the locations that we would expect, and in ways which are in line with the known dynamics of the system. The caveat to this is within the complex dynamics of the Southern Ocean. Here, although physically consistent results can be inferred, the patterns seen are complex, making it difficult to reasonably infer one particular scenario over another. It is not possible to definitively attribute increased errors to specific

https://doi.org/10.1017/eds.2022.10 Published online by Cambridge University Press

processes through this analysis, only to highlight plausible explanations, and in this complex region, multiple explanations can be inferred. This is especially notable when looking at the impacts of vertical structure in the Southern Ocean region. Here, the evidence more strongly indicates physically inconsistent inferences, indicating that the regressor has struggled to learn the full dynamics of this region. Nevertheless, it is reassuring that in most cases, and especially when looking at the north of the domain where the dynamics are less complex, physically consistent interpretations can be seen.

Our results highlight the need to perform model interpretation through a variety of methods, assessing both feature importance within models; which features are most heavily used or needed for predictive skill, and feature sensitivity; how features impact predictions. In general, we see that the regressor developed here both uses and depends on variables that are in line with the known dynamics of the system, and these variables impact predictions in ways which are consistent with our physical knowledge. These results imply that the regression model developed here is, to a large extent, learning the underlying dynamics of the system being modeled. This result is very promising in the context of further development of data-driven models for weather and climate prediction, for both atmospheric and oceanic systems.

That we see this behavior in a simple model suggests that more complex models, capable of capturing the full higher-order nonlinearity inherent in GCMs, are well placed to learn the underlying dynamics of these systems. The model developed here has a number of limitations, and a similar assessment of a more complex model, particularly one which can better capture the extreme behavior alongside the more dominant dynamics would be of value to confirm this. The work carried out here uses a very idealized and coarse resolution simulator to create the dataset used for training and validation. Further investigation into how the complexity of the training data and the resolution of the GCM used to create this dataset impacts the sensitivity of data-driven models would also be of further interest. Similarly, we assess model performance and model sensitivity over a single predictive step, but in forecasting applications, data-driven models would most likely be used iteratively. Assessment of how model skill varies when iterating data-driven models has been carried out in the context of alternative data-driven models. Looking alongside this to how the sensitivity of the model changes when using models iteratively would provide further interesting insight into this area.

As data-driven models become competitive alternatives to physics-driven GCMs, it is imperative to continue to investigate the sensitivity of these models, ensuring that we have a good understanding of how these models are working and when it is valid to rely on them.

Abbreviations

ACC	Antarctic circumpolar current
GCMs	general circulation models
GM	Gent–McWilliams
MITgcm	Massachusetts Institute of Technology general circulation model
ML	machine learning
MOC	meridional overturning circulation
RMS	root mean square
SSH	sea surface height

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

Data Availability Statement. Code used for this work (analyzing the MITgcm dataset, and training and analyzing the regression models) can be found at https://doi.org/10.5281/zenodo.5482520. The MITgcm dataset used is available at https://zenodo.org/record/4672260 (doi: 10.5281/zenodo.4672260). The generation of this dataset is described in the paper.

Author Contributions. Conceptualization: all authors; Data curation: R.F. and D.M.; Data visualization: R.F.; Investigation: R.F.; Methodology: all authors; Software: R.F.; Supervision: P.H., D.M., B.P., D.C.J., and E.S.; Writing—original draft: R.F.; Writing—reviewing and editing: all authors. All authors approved the final submitted draft. Many thanks to reviewers of this paper for their useful suggestions and the subsequent improvements.

https://doi.org/10.1017/eds.2022.10 Published online by Cambridge University Press

Funding Statement. R.F. was supported by the UK Natural Environment Research Council (grant number NE/L002507/1). D.M. was supported by the UK Natural Environment Research Council (ORCHESTRA; grant number NE/N018095/1). D.C.J. was supported by a UKRI Future Leaders Fellowship (grant number MR/T020822/1).

Competing Interests. The authors declare no competing interests exist.

References

Arcomano T, **Szunyogh I**, **Pathak J**, **Wikner A**, **Hunt BR and Ott E** (2020) A machine learning-based global atmospheric forecast model. *Geophysical Research Letters 47*(9), e2020GL087776.

Barnes EA and Barnes RJ (2021) Controlled abstention neural networks for identifying skillful predictions for regression problems. *Journal of Advances in Modeling Earth Systems 13*(12), e2021MS002573.

Breen PG, **Foley CN**, **Boekholt T and Zwart SP** (2020) Newton versus the machine: Solving the chaotic three-body problem using deep neural networks. *Monthly Notices of the Royal Astronomical Society 494*(2), 2465–2470.

Breiman L (2001) Random forests. *Machine Learning 45*, 5–32.

Chattopadhyay, A., **Hassanzadeh, P.**, & **Subramanian, D.** (2020). Data-driven predictions of a multiscale Lorenz 96 chaotic system using machine-learning methods: reservoir computing, artificial neural network, and long short-term memory network. *Nonlinear Processes in Geophysics*, *27*(3), 373–389.

Doan NAK, **Polifke W and Magri L** (2019) Physics-informed echo state networks for chaotic systems forecasting. In Computational Science (ICCS 2019), Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. *11539*. Cham: Springer, pp. 192–198.

Dueben PD and Bauer P (2018) Challenges and design choices for global weather and climate models based on machine learning. *Geoscientific Model Development 11*, 3999–4009.

Friedman JH (2001) Greedy function approximation: A gradient boosting machine. *Annals of Statistics 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Gent PR and Mcwilliams JC (1990) Isopycnal mixing in ocean circulation models. *Journal of Physical Oceanography 20*(1), 150–155.

Hoerl AE and Kennard RW (1970) Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*(1), 55–67.

Lipton ZC (2018) The mythos of model interpretability. *Queue 16*(3), 31–57.

Lorenz EN (2006) Predictability—A problem partly solved. In *Predictability of Weather and Climate*. Cambridge: Cambridge University Press, pp. 40–58.

Louppe G, **Wehenkel L**, **Sutera A and Geurts P** (2013) Understanding variable importances in forests of randomized trees. In Weinberger, C. B., Bottou, L., Welling, M., Ghahramani, Z., and K.Q., editors, *Advances in Neural Information Processing Systems*, page Volume *26*. Curran Associates, Inc.

Marshall J, **Adcroft A**, **Campin JM**, **Hill C and White A** (2004) Atmosphere–ocean modeling exploiting fluid isomorphisms. *Monthly Weather Review 132*(12), 2882–2894.

Marshall J, **Adcroft A**, **Hill C**, **Perelman L and Heisey C** (1997a) A finite-volume, incompressible Navier–Stokes model for studies of the ocean on parallel computers. *Journal of Geophysical Research: Oceans 102*(C3), 5753–5766.

Marshall J, **Hill C**, **Perelman L and Adcroft A** (1997b) Hydrostatic, quasi-hydrostatic, and nonhydrostatic ocean modeling. *Journal of Geophysical Research: Oceans 102*(C3), 5733–5752.

McGovern A, **Ebert-Uphoff I**, **Gagne DJ and Bostrom A** (2022) Why we need to focus on developing ethical, responsible, and trustworthy artificial intelligence approaches for environmental science. *Environmental Data Science 1*, E6, 1–15.

Mcgovern A, **Gagne II DJ, DJ and Lagerquist RA** (2020) Using machine learning and model Interpretation and visualization techniques to gain physical insights in atmospheric science. In *Proceedings of the International Conference on Learning Representations 2020*, 1–12.

McGovern A, **Lagerquist R**, **Gagne DJ**, **Jergensen GE**, **Elmore KL**, **Homeyer CR and Smith T** (2019) Making the black box more transparent: Understanding the physical implications of machine learning. *Bulletin of the American Meteorological Society 100*(11), 2175–2199.

Miyanawala, T. P., & **Jaiman, R. K.** (2017). An efficient deep learning technique for the Navier-Stokes equations: Application to unsteady wake flow dynamics. arXiv preprint arXiv:1710.09099.

Molnar C, **Casalicchio G and Bischl B** (2020) Interpretable machine learning—A brief history, state-of-the-art and challenges. In *ECML PKDD 2020 Workshops. ECML PKDD 2020*. Communications in Computer and Information Science, Vol. *1323*. Cham: Springer, pp. 417–431.

Munday DR, **Johnson HL and Marshall DP** (2014) Impacts and effects of mesoscale ocean eddies on ocean carbon storage and atmospheric pCO_2. *Global Biogeochemical Cycles 28*(8), 877–896.

Munday DR, **Johnson HL**, **Marshall DP**, **Munday DR**, **Johnson HL and Marshall DP** (2013) Eddy saturation of equilibrated circumpolar currents. *Journal of Physical Oceanography 43*(3), 507–532.

Pathak J, **Hunt B**, **Girvan M**, **Lu Z and Ott E** (2018) Model-free prediction of large spatiotemporally chaotic systems from data: A reservoir computing approach. *Physical Review Letters 120*(2), 024102.

Rasp S, **Dueben PD**, **Scher S**, **Weyn JA**, **Mouatadid S and Thuerey N** (2020) WeatherBench: A benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems 12*(11), e2020MS002203.

Rasp S and Thuerey N (2021) Data-driven medium-range weather prediction with a Resnet pretrained on climate simulations: A new model for WeatherBench. *Journal of Advances in Modeling Earth Systems 13*, e2020MS002405.

Rintoul SR (2018) The global influence of localized dynamics in the Southern Ocean. *Nature 558*(7709), 209–218.

Scher S (2018) Toward data-driven weather and climate forecasting: Approximating a simple general circulation model with deep learning. *Geophysical Research Letters 45*(22), 12616–12622.

Scher S and Messori G (2019a) Generalization properties of feed-forward neural networks trained on Lorenz systems. *Nonlinear Processes in Geophysics 26*, 381–399.

Scher S and Messori G (2019b) Weather and climate forecasting with neural networks: Using general circulation models (GCMs) with different complexity as a study ground. *Geoscientific Model Development 12*(7), 2797–2809.

Simonyan K, **Vedaldi A and Zisserman A** (2013) Deep inside convolutional networks: Visualising image classification models and saliency maps. In *2nd International Conference on Learning Representations, ICLR 2014—Workshop Track Proceedings*.

Stracuzzi DJ and Utgoff PE (2004) Randomized variable elimination. *Journal of Machine Learning Research 5*, 1331–1362.

Talley L (2013) Closure of the global overturning circulation through the Indian, Pacific, and southern oceans: Schematics and transports. *Oceanography 26*(1), 80–97.

Weyn JA, **Durran DR and Caruana R** (2019) Can machines learn to predict weather? Using deep learning to predict gridded 500-hPa geopotential height from historical weather data. *Journal of Advances in Modeling Earth Systems 11*(8), 2680–2693.

Appendix A: GCM Fluxes

We calculated temporally averaged advective and diffusive transports of temperature to identify which processes dominate temperature change in different regions of the domain. Figures A1 and A2 show cross sections of these transports. These are created using the same data as used in Figures 3, 6, and 7. They show an average of 500 daily transport tendencies, taken from the 20-year model dataset described previously, subsampled to take the average of every 14th day.

From these, we see that the majority of temperature change from all processes is located in the Southern Ocean and the very north of the domain. In particular, we see that the vertical advection is largest in the very north, and increased at the edge of the Southern Ocean. There is notable zonal advection of temperature around 40°S, in keeping with the high wind stress and interaction with the end of the land feature giving rise to a Southern Ocean jet—an ACC-like feature. Diffusive fluxes are generally lower (by one or two orders of magnitude). These show a broader spatial spread, although vertical, zonal, and meridional diffusive fluxes are still highest in the Southern Ocean and near the north of the domain. There is a large signal of convectively driven temperature change, due to the surface cooling in this area (applied through surface restoring). Similarly, we see increased vertical diffusive fluxes, both implicit and explicit, in the south of the domain; this is a region of high vertical activity, with both upwelling and downwelling of water masses. In the south, we also see a signal of strong meridional diffusion related to the ACC-like feature in the GCM.

Appendix B: Full Set of Coefficients

Section 4.1 and Figure 4a focused on coefficients averaged over each variable. Here, for completeness, we show the coefficients for all 26,106 inputs (Figure B1). The top row shows coefficients for the linear features (the first term in equation (1)), and the triangular lower section shows coefficients for the nonlinear terms (the multiplicative terms from the second term in equation (1)). Note that for most variables, there are 27 pixels, once for each point in the 3D neighborhood stencil.

As in Figure 4, we see the importance of density (as a linear term), the relative unimportance of location information, and the notable pattern in the multiplicative interaction between temperatures at different relative locations. Looking here at the full set of coefficients without any averaging applied, we also note that the interaction of the U component of the current with temperature shows one or two points which are highly weighted. We suspect that this is symptomatic of regression models trained by least-squared error methods being highly sensitive to a few extreme training samples. When developing the regressor, we saw that different versions of the model all had a tendency to weight one or two coefficients very highly, with each iteration of the model favoring different coefficients. The more general patterns, particularly those seen when averaging over variables, and the sharp stripes in the temperature–temperature interactions, were consistent through all versions of the regressor, just the occasional very small number of highly weighted inputs changed. This indicates a lack of robustness in the model, meaning that small changes to the training dataset can erroneously cause a few terms in the equation to become very highly weighted. This highlights the need to consider the robustness of data-driven models, and their sensitivity to the training samples used, and also emphasizes the importance of using more than one technique to assess feature importance.

Appendix C: Predicting over Longer Timescales

We ran three additional versions of the regression model predicting 5, 10, and 20 days ahead, and compared the results with the regressor predicting a single forecast day ahead. To clarify, this was not based on using the regressor iteratively, as the regressor is not designed to be used in this way. Instead, the regressor makes a single forecast step of 5, 10, or 20 days, in place of the 1-day forecast

https://doi.org/10.1017/eds.2022.10 Published online by Cambridge University Press

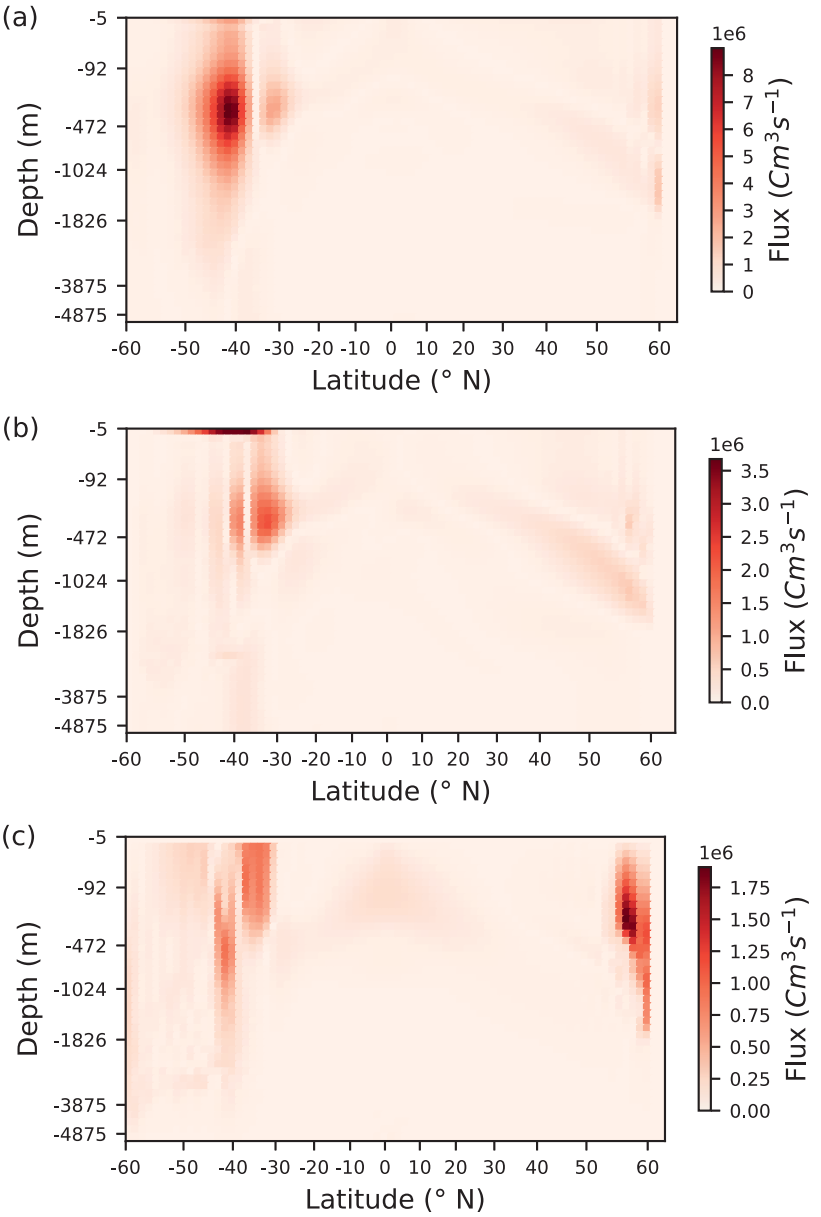

Figure A1. *Average absolute zonal (a), meridional (b), and vertical (c) advective fluxes of temperature at 13° E. Horizontal advective fluxes are largest in the southern region of the domain, associated with the ACC-like current. There is a large amount of vertical advection in the north of 55°, and at −30 to −40°, associated with regions of upwelling and downwelling.*

step used in the control and throughout the paper. We consider the effect this has on the predictions. Table C1 shows the root-mean-square (RMS) error and skill score for the regressors trained to predict 1, 5, 10, and 20 days ahead, along with the RMS errors for persistence forecasts over the same forecast length.

We can see that the RMS errors grow larger with longer forecast lengths, over both the training and validation sets, meaning that predictions have greater error over longer forecast lengths. This is to be expected, as predicting further ahead is a more challenging task. Temperature changes are larger over longer time periods, and the dynamics of the underlying simulator (and the real ocean) mean that the temperature change at a particular point over a longer time period is driven by points increasingly further away, and in increasingly nonlinear ways. As we only provide the regressor with information from directly neighboring points as inputs, when

https://doi.org/10.1017/eds.2022.10 Published online by Cambridge University Press

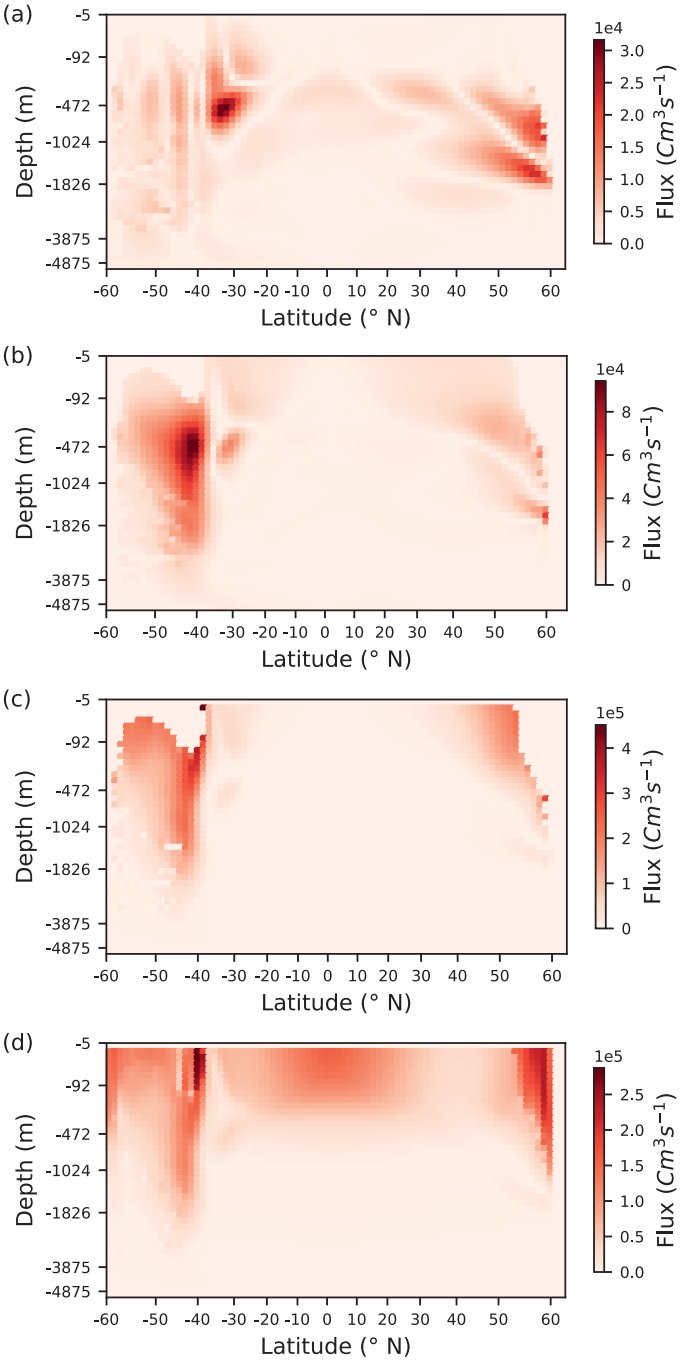

Figure A2. *Average absolute zonal (a), meridional (b), and (explicit) vertical (c) diffusive temperature fluxes, and convective (implicit vertical diffusive) temperature fluxes (d) at 13° E. There are large amounts of meridional diffusion associated with the ACC-like jet in the south. Zonal diffusion occurs in mid depth in the north of the domain, and just north of −40°. Vertical diffusion occurs through the south of the domain, and a small region just south of 50°. Convection occurs throughout the domain, and is particularly noteworthy in the upper waters of the ocean north of 50°, and south of −35°.*

https://doi.org/10.1017/eds.2022.10 Published online by Cambridge University Press

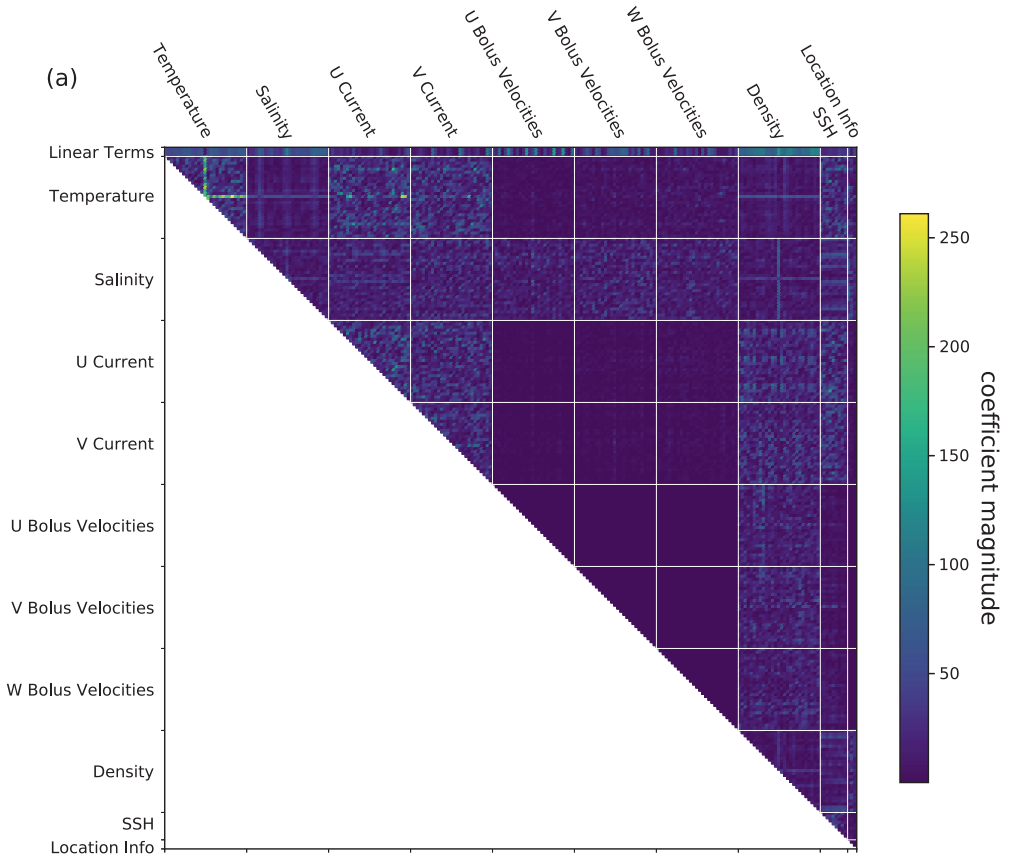

Figure B1. *Coefficients of the control regressor for each input location and for each variable type. For linear inputs (top row) and for each set of nonlinear combinations of variables.*

Table C1. Table showing RMS errors and skill scores for four models trained to predict temperature change over an increasing forecast period. RMS errors increase with forecast period, but skill scores are largely unaffected.

	RMS error (°C)		Skill score $\left(1-\frac{modelRMS}{PersistenceRMS}\right)$	
	Training	Validation	Training	Validation
Control (1-day forecast step)	5.61e-5	9.89e-5	0.45	0.14
Persistence over 1-day forecast step	1.02e-4	1.15e-4	—	—
5-day forecast step	2.79e-4	4.72e-4	0.45	0.14
Persistence over 5-day forecast step	5.07e-4	5.49e-4	—	—
10-day forecast step	5.56e-4	9.27e-4	0.45	0.14
Persistence over 10-day forecast step	1.00e-3	1.08e-3	—	—
20-day forecast step	1.07e-3	1.83e-3	0.45	0.14
Persistence over 20-day forecast step	1.95e-3	2.13e-3	—	—

Abbreviation: RMS, root mean square.

https://doi.org/10.1017/eds.2022.10 Published online by Cambridge University Press

looking at temperature changes over longer time periods, when points further away influence temperature change, the regressor is increasingly limited by the lack of input information. Similarly, as the regressor is only able to represent a small amount of nonlinearity, we would expect predicting further ahead to become more challenging.

We consider how much of this increased error is related to the problem becoming harder with longer forecast step, or if there is any indication that the regression model is inherently unsuitable for forecasting over these longer forecast steps. By incorporating the baseline persistence RMS error, which also increases as the problem becomes harder, the skill score gives an indication of this differentiation. We see that the skill scores remain constant (to two significant figures) regardless of the length of forecast step. This shows that while the model RMS error increases, this is likely to be due to the increasing difficulty of the prediction problem, and not a sign that the model itself is unsuited to predicting across these longer timescales.

This is a particularly interesting result in the context of data-driven forecasting. Traditional GCMs, such as the MITgcm simulator used to create the training and validation datasets, are limited in the length of forecast step that can be taken due to numerical constraints. At some point, a GCM would show large numerical errors due to numerical instabilities, alongside the expected growth in errors related to the increased difficulty of the prediction problem. For the configuration shown here, however, we obtain similar skill scores with a data-driven model when forecasting over far larger steps than would be possible in the simulator. This indicates that data-driven models are more stable when predicting over long time periods, meaning that if suitable inputs were provided to enable accurate results over long time periods, this type of model could be far more efficient than traditional GCMs, particularly for climate runs. These results warrant further investigation, in particular to see if similar patterns are shown with more complex configurations. It would also be of interest to investigate whether the sensitivity of the regressor changes with increasing forecast length.

Cite this article: Furner R. Haynes P. Munday D. Paige B. Jones DC and Shuckburgh E. (2022). A sensitivity analysis of a regression model of ocean temperature. *Environmental Data Science*, 1: e11. doi:10.1017/eds.2022.10