Reviewer 2

Many thanks for your helpful comments and suggestions, and for taking the time to read and review our paper. We have responded to the comments in blue below (with the reviewers original comments in grey). Where we have made changes to the manuscript we've noted these below our comments, with red strikethrough text for deletions, green text for additions and black text indicating regions of no change.

Recently, the data-driven models have become a hot topic for the atmospheric and oceanic predictions for time scale from synoptic to interannual. Many cases indicate that such models have high predict skills and much less computational resource-consuming. However, as mentioned by the authors, some fundamental questions have no been answered yet, such as if the method can capture the predictability natural of the system and if we can physically explain the results. In the present study, a data-driven model has been developed for predicting the sea surface temperature in the short term. The method they used is a regression model with a non-linear term. The model has been trained using an idealized ocean model dataset as an observing system simulation experiment. They found that the model can predict the SST leading one day, and the dominant variables are also identified. After that, the sensitivity withholding experiments are conducted to identify critical physical variables and processes, like the vertical structure and the non-linear term. In general, this is an interesting and valuable paper and provides helpful information for this kind of data-driven model. Therefore, it is worth to published in GMD after the MINOR revision.

Just to note that the model predicts sea temperature at a range of depths - it is not intended to predict the change in SST, but rather it predicts the change in temperature at depth throughout the interior of the ocean.

Major questions:

1. I think the most critical problem is the resolution of the model is too coarse (2 degrees) compared with the predicting time scale (1 day). Because the movement of the ocean is much slower than that of the atmosphere, the surrounding points cannot affect the central point for one day. The only fast process is the convection due to instability. That is why the coefficient magnitude of the center point is much larger than other points in Figure 4. If the horizontal resolution is increasing, I guess that the results may also be changed because some processes may transport the signal of around points to the central point during one day. Therefore, I suggest the authors test the sensitivity of the resolution further.

Firstly, we stress that figure 4 does not show the impact of the central point alone. The top panel (figure 4a) is the impact averaged over all input points (the central point and all neighbouring point), and the bottom panel (figure 4b) shows the interaction between pairs of input points. Assuming the bright banding in figure 4b is what is being referred to here, this shows the polynomial terms are heavily weighted which combine information from the central point and surrounding points. This reflects the importance of *interaction* between the temperature at the central point and its neighbours, rather than the importance of the central point itself. We would expect this near-neighbour interaction to remain a dominant feature at all resolutions. We've amended the caption for figure 4 to clarify this;

Figure 4 caption:

> Coefficients of the control regressor. Top: Coefficients averaged over all input locations for each variable type, and each set of non-linear combinations of variables. Bottom: ~~Coefficients for Temperature-Temperature interactions, with values for each input location and for each interaction between input locations shown~~ Coefficients for

> polynomial terms representing Temperature-Temperature interactions across all pairs of input locations.

Regarding testing the sensitivity of the resolution, thanks for this suggestion - it is certainly an interesting question as to how the resolution would change the ability of data-driven methods to learn the dynamics of the ocean and the sensitivity of the learned equations.

To a large extent these questions around the interaction between the spatial resolution used and the predicting time scale would be equally well considered by changing the temporal resolution - something we have considered in response to point 5, please see below.

Unfortunately assessing the sensitivity of the model to varying spatial resolution would require significant effort (it would involve re-running multiple simulations of the underlying MITgcm simulator, training the full set of new regressors on this data, and then analysing these results). It would also take considerable computational resource in comparison to the existing work - even doubling the resolution would take 10 times the compute resource running the simulator and would result in around 4 times as much data, which in turn also lead to additional compute resource needed to train and validate the regressor.

Whilst acknowledging this investigation would make for interesting further work, we do not feel this is in the scope of this paper. We've added the following to the manuscript to reflect the potential of this for further work.

~Line 513:

> That we see this behaviour in a simple model suggests that more complex models, capable of capturing the full higher-order non-linearity 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 behaviour alongside the more dominant dynamics would be of value to confirm this. The work carried out here uses a very idealised 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, impact 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 also provide further 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 we have a good understanding of how these models are working and when it is valid to rely on them.

2. In the withholding experiments, we can find that the errors are smaller than the control experiments in some places, such as withholding the information about the vertical structure in Figure 6 and withholding currents in Figure 7. How to understand these results? Is there some information that will bring negative effects or significant errors into the model? Can we find an optimal combination of all this information?

We have made some changes to the manuscript to further clarify this point. We hope this clarifies things. It is not the case that the additional information brings errors to the model, but that the best *overall* equation for the control is not better at every individual prediction, compared to the best *overall* equation developed in each of the withholding experiments.

~Line 408:

> 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 2-d information. ~~Given the regressor we have developed is learning one equation to be applied across all grid boxes in the domain it is not unexpected that removing variables might improve performance in a few small regions if the restricted equation is a better fit for that particular location, but not more favourable across the entire domain.~~ In this work we have developed a regressor which learns one equation to be applied across all grid boxes in the domain. We optimise 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 versions of the model will favour certain types of dynamics than others. Given this, it is not unexpected that the new equations discovered for the withholding experiments (which also optimise 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 optimises for good performance overall (see fFig. 3b), rather than the much more varied performance seen in the withholding experiments (Fig. 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 optimises performance in other regions to minimise error overall. This highlights the limitations of our method, and the need for more complex data-driven models which 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 are optimised for different locations within the domain, and combine these to produce a domain wide 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, would 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 methods could be explored however to find a single model which has the complexity to better capture the detailed non-linear dynamics in the ocean.

~Line 447:

> 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 models. This is for the same reasons as described in Sect 4.3.2

3. The configurations of the model or experiment are not present very clear. Please give more information about the experiment in the manuscript, such as the vertical levels, the time scale of the restoring, the observation of the restoring sea surface temperature and salinity, and the coefficients of the GM scheme.

Thanks for this suggestion. The underlying MITgcm simulation is published in Munday et al (2013)[1], but for reader ease we have expanded the description of the configuration within the

[1] Munday, D. R., Johnson, H. L., Marshall, D. P., Munday, D. R., Johnson, H. L., and Marshall, D. P.: Eddy Saturation of Equilibrated Circumpolar Currents, Journal of Physical Oceanography, 43, 507–532, https://doi.org/10.1175/JPO-D-12-095.1, http://journals.ametsoc.org/doi/abs/10.1175/JPO-D-12-095.1, 2013

paper, including adding a table with key parameters, and pointed the reader more clearly to the aforementioned paper.

~Line 95:

This configuration, whilst 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).

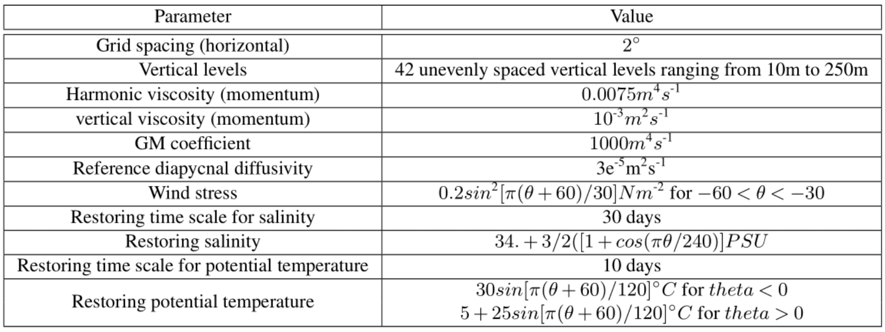
The domain runs from -60°S to 60°N, and is just over 20° wide in longitude. The domain is bounded by land along the northern (and southern) edge and a strip of land runs along the eastern (and western) boundary from 60°N to 40°S (see Fig.1a). Below this, in the southern-most 20°, the simulator has a periodic boundary condition, allowing flow which exits to the east (west) to return to the domain at the western (eastern) boundary. The domain has flat-bottom bathymetry of 5000m over most of the domain, with a 2° region of 2500m depth at the southern-most 20° of the eastern edge (i.e. the spit of land forming the eastern boundary continues to the southern boundary as a 2500m high sub surface ridge)

The simulator is run with 42 (unevenly spaced) depth levels, following a Z-coordinate, with the surface layer being the thinnest at 10m, and the bottom 10 levels being the maximum at 250m ~~, and has~~ There are 11 cells in the X (longitudinal) direction, and 78 cells in the Y (latitudinal) direction. The grid spacing is 2° in the Y direction, with the X spacing scaled by the cosine of latitude to maintain approximately square grid boxes (this means 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 hour time step (two steps per day), with fields output daily. We focus on daily-mean outputs, rather than the instantaneous state.

At 2° resolution the simulator is not eddy resolving, but uses the Gent–McWilliams parameterisation (Gent and Mcwilliams, 1990) to represent the effects of ocean eddy transport. We ran the simulator with a strong surface restoring condition (see Table 1) on Temperature and Salinity — thus fixing the surface density. We apply 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 Nm^{-2} at 45°S.

Table 1:

Table 1. Key parameter information for MITgcm simulation



Parameter	Value
Grid spacing (horizontal)	2°
Vertical levels	42 unevenly spaced vertical levels ranging from 10m to 250m
Harmonic viscosity (momentum)	$0.0075 m^4 s^{-1}$
vertical viscosity (momentum)	$10^{-3} m^2 s^{-1}$
GM coefficient	$1000 m^4 s^{-1}$
Reference diapycnal diffusivity	$3e^{-5} m^2 s^{-1}$
Wind stress	$0.2 sin^2[\pi(\theta+60)/30] Nm^{-2}$ for $-60 < \theta < -30$
Restoring time scale for salinity	30 days
Restoring salinity	$34. + 3/2([1 + cos(\pi\theta/240)] PSU$
Restoring time scale for potential temperature	10 days
Restoring potential temperature	$30 sin[\pi(\theta+60)/120]^\circ C$ for $theta < 0$ $5 + 25 sin[\pi(\theta+60)/120]^\circ C$ for $theta > 0$

4. The method of selecting data is also not present very clear. For instance, I do not really understand how to choose the data every 200-day and deal with the 3D variables at the surface. Please say something more about the details.

Thank you for highlighting the need for more explanation here. We have made the following changes to the text to clarify these points.

~Line 136:

~~As the data is highly auto-correlated we sub-sample in time to remove some of the co-dependent nature of the training data. There are also computational constraints limiting the total size of our dataset. This leads us to choose a subsampling rate of every 200th day.~~ The data is highly auto-correlated, i.e. fields are similar, particularly when considering fields which are temporally close. This strong auto-correlation, found in many weather and climate applications, impacts the ability of the algorithm. Therefore, as is common practice, we sub-sample in time to remove some of the co-dependent nature of the training data, better optimising 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), whilst also fitting within computational constraints, and limiting auto-correlation within the dataset.

~Line 140:

~~For every 200th day, we take all gridpoints from throughout the model interior (i.e. we exclude points next the land, and points at the surface and seabed).~~ For every 200th day, 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.

5. In the present study, the authors only show the results of one-day prediction. I am curious how the model performs when the predicted time scale becomes longer, like 5-day or 10-day. I suggest the authors further discuss the skill of the model for more extended timescale prediction.

Many thanks for this suggestion. We have added a short appendix (Appendix C in the manuscript) showing how the model performs when predicting over a 5-day, 10-day and 20-day period, and a brief discussion of these results.

Appendix C: Predicting over longer timescales

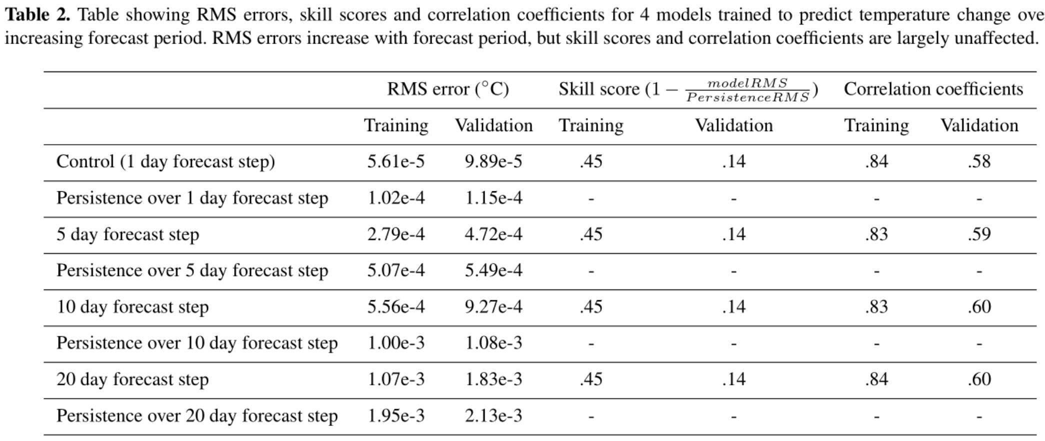


Table 2. Table showing RMS errors, skill scores and correlation coefficients for 4 models trained to predict temperature change over increasing forecast period. RMS errors increase with forecast period, but skill scores and correlation coefficients are largely unaffected.

	RMS error (°C)		Skill score ($1-\frac{modelRMS}{PersistenceRMS}$)		Correlation coefficients	
	Training	Validation	Training	Validation	Training	Validation
Control (1 day forecast step)	5.61e-5	9.89e-5	.45	.14	.84	.58
Persistence over 1 day forecast step	1.02e-4	1.15e-4	-	-	-	-
5 day forecast step	2.79e-4	4.72e-4	.45	.14	.83	.59
Persistence over 5 day forecast step	5.07e-4	5.49e-4	-	-	-	-
10 day forecast step	5.56e-4	9.27e-4	.45	.14	.83	.60
Persistence over 10 day forecast step	1.00e-3	1.08e-3	-	-	-	-
20 day forecast step	1.07e-3	1.83e-3	.45	.14	.84	.60
Persistence over 20 day forecast step	1.95e-3	2.13e-3	-	-	-	-

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 day 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 step used in the control and throughout the paper. We consider the effect this has on the predictions. Table C1 shows the RMS error, skill score and correlation coefficients 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 correlation coefficient changes very little between the regressors. Correlation coefficients indicate how well the regressor captures the ’pattern’ of the data, and at all forecast lengths the regressors do well at capturing these.

Generally RMS error is a more useful statistic in forecasting problems, as it gives an indication of average error per prediction. We can see that the RMS is larger with longer forecast lengths, over both the training and validation sets, meaning 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 non-linear ways. As we only provide the regressor with information from directly neighbouring points as inputs, when 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 non-linearity, we would expect predicting further ahead to become more challenging.

We also 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 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. For the configuration shown here however, we obtain similar skill even when forecasting over far larger steps than would be possible in the simulator, making this type of model far more efficient. These results warrant further investigation, in particular to see if similar patterns are shown with more complex configurations, and if the sensitivity of the regressor changes with increasing forecast length.