Developing a data-driven ocean model – sensitivity of a linear regressor

Rachel Furner (British Antarctic Survey and University of Cambridge), Peter Haynes (University of Cambridge), Dan Jones (British Antarctic Survey), Dave Munday (British Antarctic Survey), Brooks Paige (Alan Turing Institute & UCL), Emily Shuckburgh (University of Cambridge)

@rachelafurner
rachel.furner@maths.cam.ac.uk

Introduction

This work looks at creating a data-driven analogue model of an MITgcm configuration. In the first instance we develop a linear regressor to test the capability of simple statistical methods, and then assess sensitivity of this regressor to its inputs. This enables us to assess the importance of various inputs, and from this gain some intuition as to whether the model learns in a way which matches our understanding of the physical ocean and its dynamics. The regressor also provides a baseline for expected skill from more complex data-driven methods, such as deep neural networks,

MITgcm dataset

We take a 2^{o} resolution channel-configuration of MITgcm following Munday et al (2013). This has a single 5km deep basin, with a periodic flow in the South over a 2.5km deep ridge. We run this with constant forcing: a constant 'jet' of wind stress applied with a sine based distribution between 60^{o}S and 30^{o}S, with a peak value of 0.2 N m^{-2} s^{-1}, and strong surface restoring for Temperature and Salinity. The model runs with a 12-hour timestep, and we output daily mean values of all variables. The domain can be seen in figure 1, which shows the temperature field at a given depth (z=5), along with the same field one day later and the difference between these.

We run MITgcm for 100 years under the constant forcing, during this time period the model remains very dynamically active – i.e. it is not nearing an equilibrium state. We discard the first 50 years as model spin up – when the model is responding to inconsistent dynamics in its initial conditions, and thus behaving in a different way to that expected outside of this spin up time. This leaves us with output from a 50 year run.

Data from the this 50 year run is subsampled in time (to try to ensure samples are ~ independent), and taken from each spatial location. This datatset is split into training, validation and test data, with a 70-20-10 split, with data is systematically selected from the first 70% of the run, the next 20%, and then the next 10%, so the different datasets contain data from different temporal sections of the run, this ensures the different datasets are truly independent (otherwise data could be highly co-varying). This process gives datasets of sizes 648,440 training samples, 199,520 validation samples and 99,760 test samples.

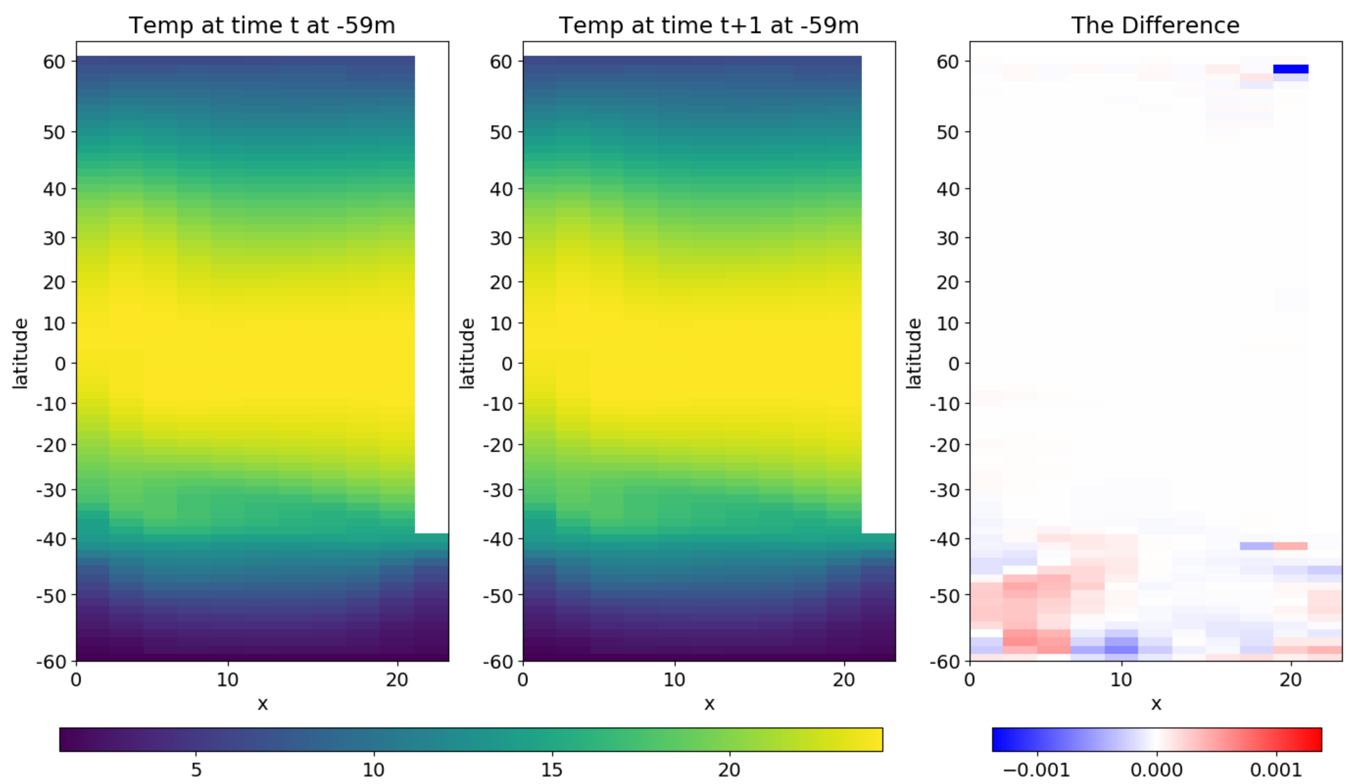

Figure 1: Daily mean temperature in ^{o}C at depth z=5, from one two consecutive days, and their difference

Inputs and Outputs

We train a linear regressor to predict change in temperature for a single grid cell. The output is therefore a single variable - the difference between daily mean temperature, at the grid point being evaluated, at two consecutive days; temp[t+1]-temp[t] where t is the time at which inputs are evaluated. i.e. this is the change in temperature between the current time step (when input information is is available) and the next timestep.

The input variables are:

- The following variables evaluated at a 3d neighbourhood of points, given by a 3x3x3 stencil centred on the grid being predicted (i.e. 27 points), as daily mean values evaluated at time t:
  - Temperature
  - Salinity
  - U&V components of current
  - The variable Kwx, Kwy and Kwz from the Gent-McWilliams tensor, used in calculating Bolus Velocities
  - Density, estimated using the simplified equation of state)
- Daily mean SSH anomaly at the 2d neighbouring points given by a 3x3 stencil (as this is a 2-d variable) evaluated at time t.
- Latitude of the grid point being evaluated
- Longitude of the grid point being evaluated
- Depth of the grid point being evaluated

This gives an original set of 228 features, from which 2nd order polynomial interactions between these variables are also included, giving 26,106 features (note square terms are not included, just interactions between features)

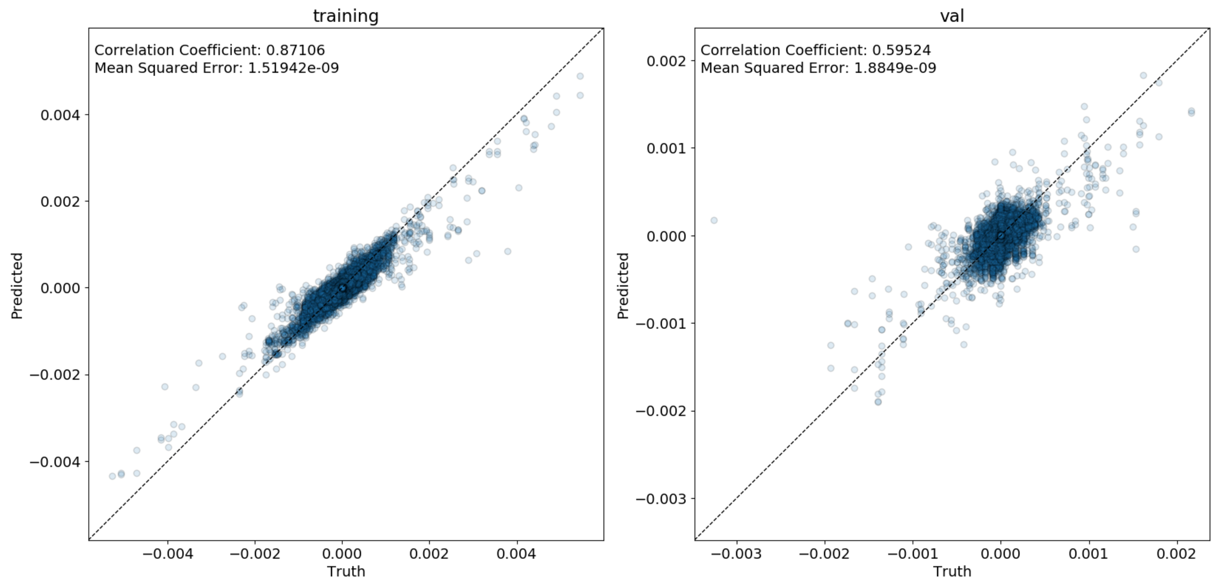

Figure 2: Scatter plot of truth (change in temperature in ^{o}C from MITgcm) against linear regressor prediction for training (l) & validation (r) sets for the control run (with all variables included)

Linear Regressor

A linear regressor Is then trained on this dataset, using ridge regression, with alpha chosen by cross validation, with values ranging from 0.0001 to 1.0. The linear regressor performance for the control (the full dataset) is shown in figure 2, and in table 1 (first row). The regressor clearly captures the behaviour of the system. As would be expected performance is better on the training set, however the drop in skill over the validation set is notable. It may be that increased regularisation is needed, given the model appears to overfit significantly.

Table one also includes results from a persistence forecast as a baseline for skill (last row). A persistence forecast is one which no change is forecast, so later time steps are identical to previous time steps. As we are forecasting the change in temp, this means all predictions would be zero and so in the scatter plots, a persistence forecast would show as all points lying on the straight horizontal line of y (predicted) = 0. From the right hand pane of figure 1 we can see that the changes we are forecasting are very small (note the scale) and confined to small regions. Over much of the domain, the change in daily mean temperature over a day is far less than 0.001, in particular, at mid to deep depths the changes are very minimal (not shown here). As such, a persistence forecast provides good predictions for this problem.

That the linear regressor exceeds these scores statistically, and is shown in figure 2 to capture the dominant dynamics of the system is impressive.

It is also worth noting that that the skill of the persistence forecast changes over the training and validation datasets. Histograms of the datasets (not presented here) show that the range of temperature change in the validation set is smaller than that in the training set – further assessment of the dataset shows this is likely an artefact of the subsampling, rather than a change in the dynamic regime of the MITgcm simulation – as the majority of samples are from times and locations when the change in temperature is small, and the smaller size of the validation set, along with the subsampling in time, means far fewer points with large temperature changes are included in the validation dataset. This does mean the distribution of the validation set differs from that of the training set, however, the underlying dynamics we are trying to replicate remain the same.

Figure 3a shows the spatial nature of errors from the regressor. We produced 1000 forecasts for each point in the domain, by running the regressor with 500 different input sets – each input set corresponding to the state output from MITgcm from one of 1000 consecutive days. These are then averaged to give an average error for each location in the domain. Figure 3 shows a North-South cross section the errors are by far largest in regions where there are more dynamics, and more change in temperature. In particular errors are large in the South of the domain, the simulated Southern Ocean, throughout depth. There are also notable errors in the North of the domain. Both these regions are locations where there are both increased dynamics generally, and increased vertical motion.

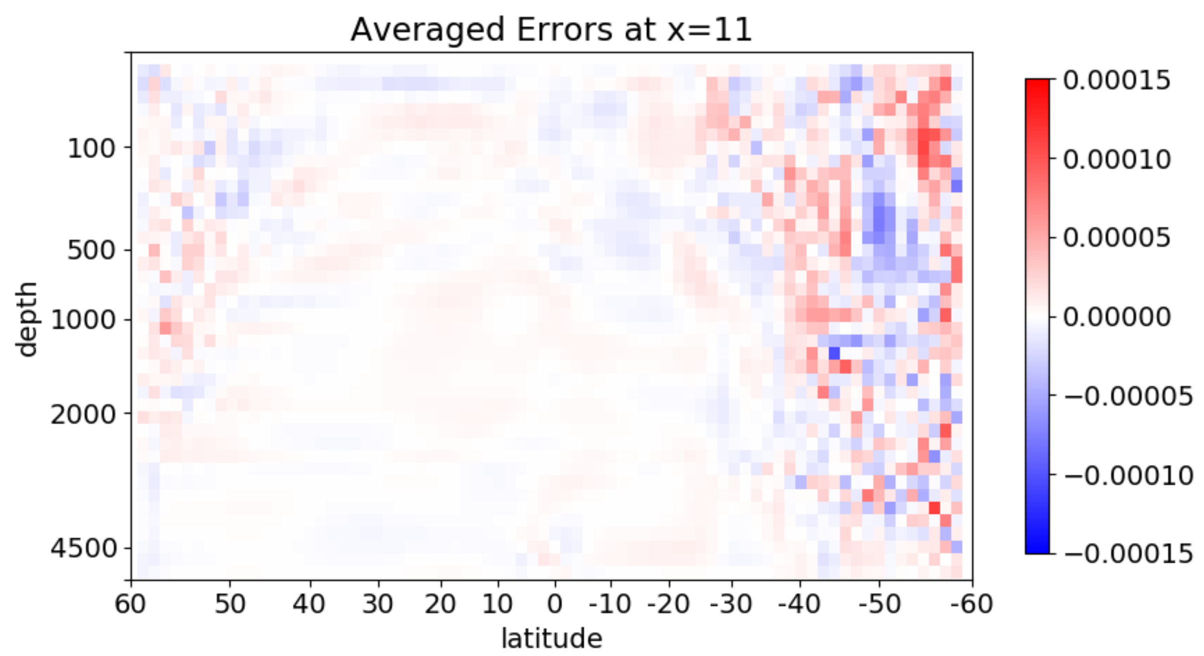

Figure 3: North-South cross section at x=5 of errors averaged over 500 one-time step predictions of the iterator, from the control (full dataset)

Sensitivity studies

A number of withholding experiments were run to assess the sensitivity of the regressor to its inputs. These experiments entail re-training the regressor with a dataset which includes all bar one of the input variables (i.e, one variable is withheld from the dataset). RMS errors from these are shown in table 1. In all cases the regressor performs better, in some cases substantially so, than a persistence forecast when assessed on the training data, indicating there is skill in the linear model for all sets of inputs tested. However, when looking at performance over the validation set, the RMS error is more comparable. Generally, the regressor still outperforms persistence, with the exception of three experiments: withholding currents, without polynomial interactions, and using only 2d information – these are discussed further below.

These sensitivity experiments highlight a number of interesting points, and insight into how the regressor is working, some of which are briefly discussed.

	Training RMS Error	Validation RMS Error
Control (full input set)	3.90e-05	4.34e-05
Withholding depth	3.93e-05	4.34e-05
Withholding latitude	3.93e-05	4.36e-05
Withholding longitude	3.93e-05	4.36e-05
Withholding Eta (sea surface height)	4.14e-05	4.43e-05
Withholding Salinity	4.40e-05	4.46e-05
Withholding density	4.41e-05	4.47e-05
Withholding Bolus velocity	5.56e-05	4.30e-05
Withholding Currents	5.77e-05	5.03e-05
Using a 2-d (3x3) input stencil	6.55e-05	4.93e-05
Without polynomial interactions	7.90e-05	4.80e-05
Persistence model (for comparison)	7.93e-05	4.78e-05

Table 1: Train and Validation scores for the control, a number of sensitivity tests, and a persistence forecast. Runs are ordered by training error

Lack of sensitivity to location data: Whilst withholding spatial information (latitude, longitude and depth of the grid points) does reduce the accuracy of the model, this has a limited impact on results compared to withholding other variables. This is reassuring, as given the constant forcing applied in the MITgcm configuration being used, one concern was that the regressor would find a non-dynamic correlation based primarily on location – that this is not the case, and that location parameters have the smallest impact in these experiments, indicates that the regressor is 'learning' something strongly related to the dynamics. The performance of the regressor is far more dependant on the physical ocean variables than on location information.

Importance of non-linearity: The worst performance from the regressor is when 2nd order polynomial combinations of the variables are not included, showing the regressor requires some interaction between variables in order to forecast well. This is in keeping with our understanding of the physical system, which is known to exhibit complex non-linear behaviour. Figure 4 shows scatter plots for the experiment which excludes non-linear interactions, comparison between this and figure 1 shows the impact of including polynomial terms. Without non-linearity the regressor predicts near zero change for all points, giving a forecast very similar to that of the persistence forecast. Non-linearity is essential for the regressor to capture any of the variability of the system.

Vertical Processes: The regressor also performs notably poorly when only 2-d information is provided. Figure 5 is a similar plot to figure 3, but this time from the run which uses only a 2 dimensional (3x3) stencil for inputs. Comparing this to figure 3, we can see that excluding information on the vertical structure of the physical ocean variables particularly increases errors in the far north of the domain, and in the Southern Ocean. These are both regions where vertical processes are key, as regions of intense upwelling and downwelling.

A similar increase in errors in the far north of the domain, and in the Southern Ocean region is also seen in the runs which excludes inputs related to the bolus velocities, and to a lesser extent in the run which excludes a pre-calculated density. Both these are innately related to vertical processes – The bolus velocity is related to the Gent-McWilliams mixing scheme, and density drives vertical motion in regions of instability, and inhibits vertical motion in stable regions.

As we know the dynamics in both the far north of the domain, and in the Southern Ocean regions involve considerable vertical process it is reassuring to see that these are the regions the regressor struggles with when inputs relevant to these processes are withheld.

Importance of currents: Of all the physical ocean variables tested, the currents show the largest impact on RMS scores. This is again in keeping with out physical understanding of the system – a primary driver of temperature change in the ocean is the movement of water through wind driven currents, especially in this configuration which excludes surface forcing. Without this information, the regressor is very limited in its ability to predict temperature change.

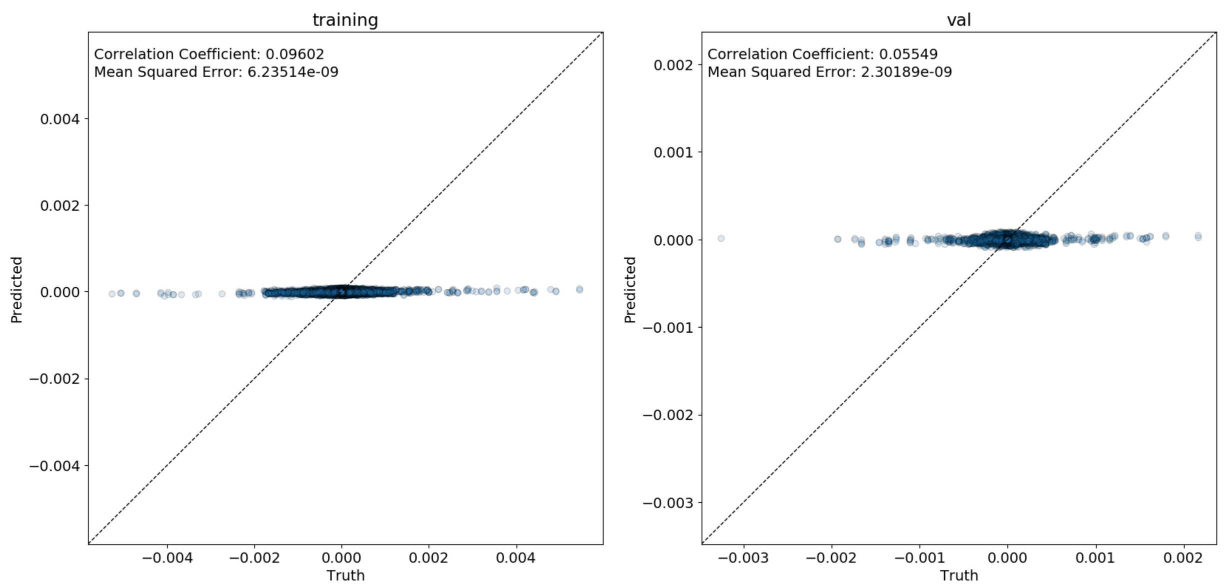

Figure 4: Scatter plot of truth (change in temperature in ^{o}C from MITgcm) against linear regressor prediction for training (l) & validation (r) sets when polynomial interactions are not included in the regression model

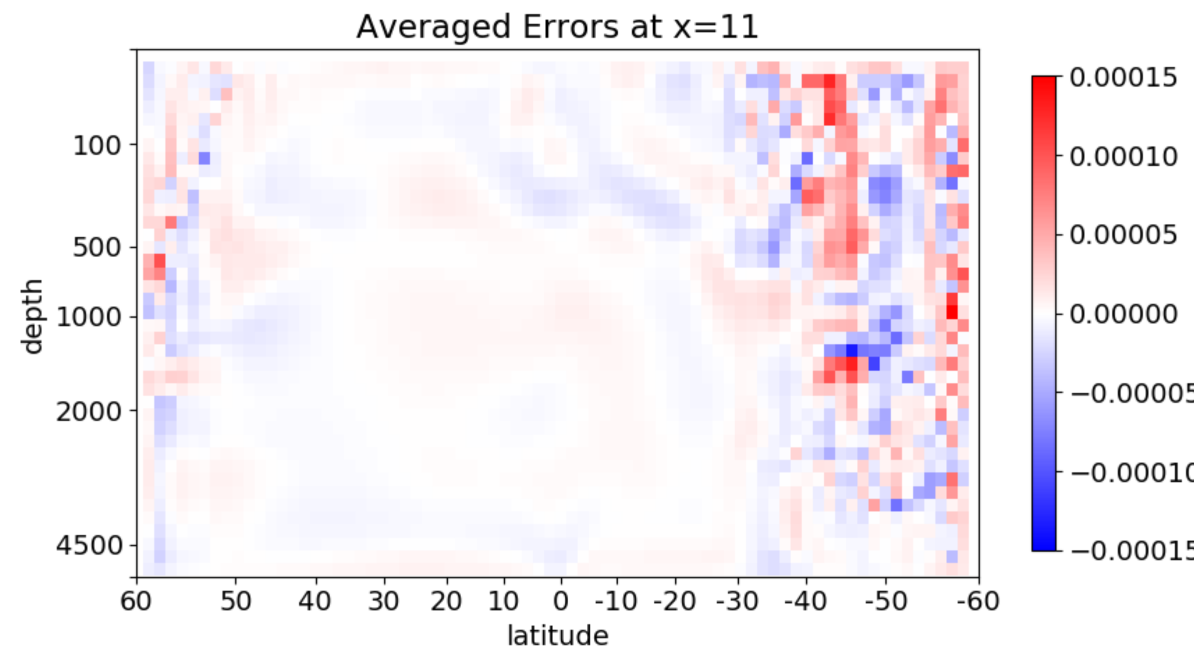

Figure 5: North-South cross section at x=5 of errors averaged over 500 one-time step predictions of the iterator, from the run with a 2-d stencil

Summary

- We've developed a linear regressor which does a good job of predicting change in temperature from one day to the next, using information from the current day as inputs.
- Sensitivity studies show the regressor behaves in a way consistent with the known dynamics of the system:
  - The regressor requires non-linear interactions in order to capture the variability of the system
  - The regressor has far higher sensitivity to physical ocean parameters than location parameters
  - Investigation of the structure of errors shows that errors are higher in areas associated with important vertical processes when the inputs related to these processes are withheld
- These findings indicate that the linear regressor has not only learnt to predict change in temperature, but that these predictions are based on the regressor having 'learnt' the underlying dynamics of the system.

Further work

- Further investigation into the formation and pattern of errors when groups of inputs (i.e. all spatial information, or all information relating to vertical processes) are withheld simultaneously would be interesting in further assessing the importance of these processes.
- The sensitivity studies from the regressor indicate that capturing non-linearities in the system is key to making good predictions, but the regressor is still limited by the extent to which it can capture non-linearities. More sophisticated methods are needed, as such, we've begun preliminary studies with various network approaches to this problem and plan to continue this.
- To date, we have treated the problem as a Markov Chain when clearly this is not an accurate assumption. We plan to assess the impact of this, by including more history to the regressor (so inputs come from t-1 as well as t), and to look into more sophisticated statistical and machine learning methods, such as LSTM models, and echo state networks.