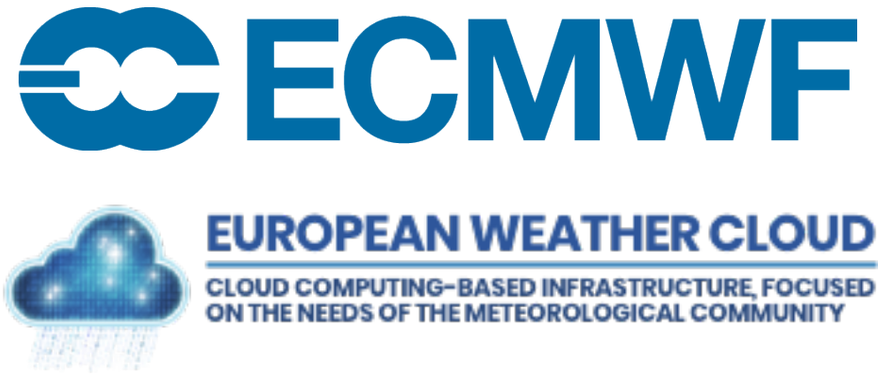
Emulation of gravity wave parameterisation in weather forecasting

Matthew Chantry[1], Sam Hatfield[2]
Peter Düben[2], Tim Palmer[1]
[1]University of Oxford, [2]ECMWF



What are parameterisation schemes?

- One of many kernels of a weather or climate model.

- Each scheme captures an aspect of unresolved physics.

- Assumptions/approximations/observations are used to create closed schemes.

- Act on vertical column of data, produce an increment for variables in the column.

Our questions

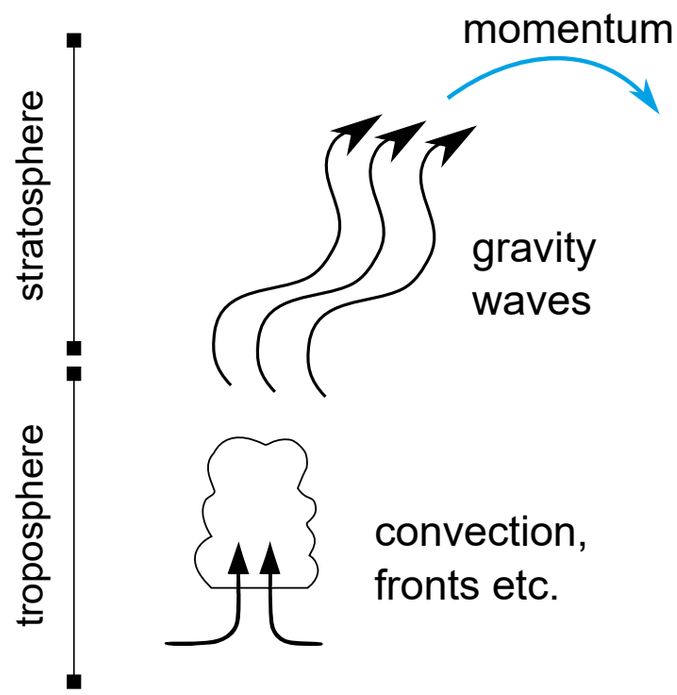
- Can we emulate parameterisation schemes with neural networks?

- Are the emulators cheaper than the originals?

- Can this help with data assimilation?

- Do lessons transfer between schemes? e.g. types of networks, normalisation methods etc

Which schemes?



- Schemes with IFS code.

- <u>Orographic</u> and <u>non-orographic</u> gravity wave drag.
  - Captures the impact of unresolved gravity waves on resolved scales.

  - Take u, v ,T and description of the model levels, produce increments for u, v, T. (Additional variables for orography).

  - Why? Simple schemes, capturing similar processes but with key differences (see later)

Data

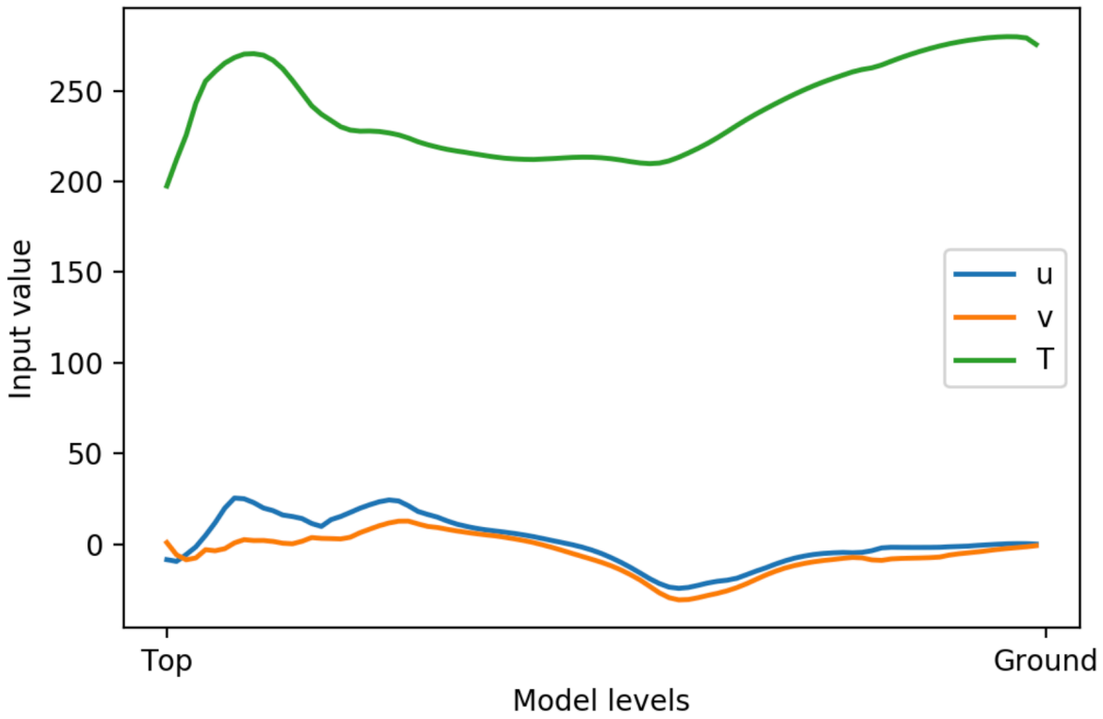
- Simulate 2 years in IFS (TCo399 ~18km grid).

- 2,355,840 data pairs per month (every 5hrs).

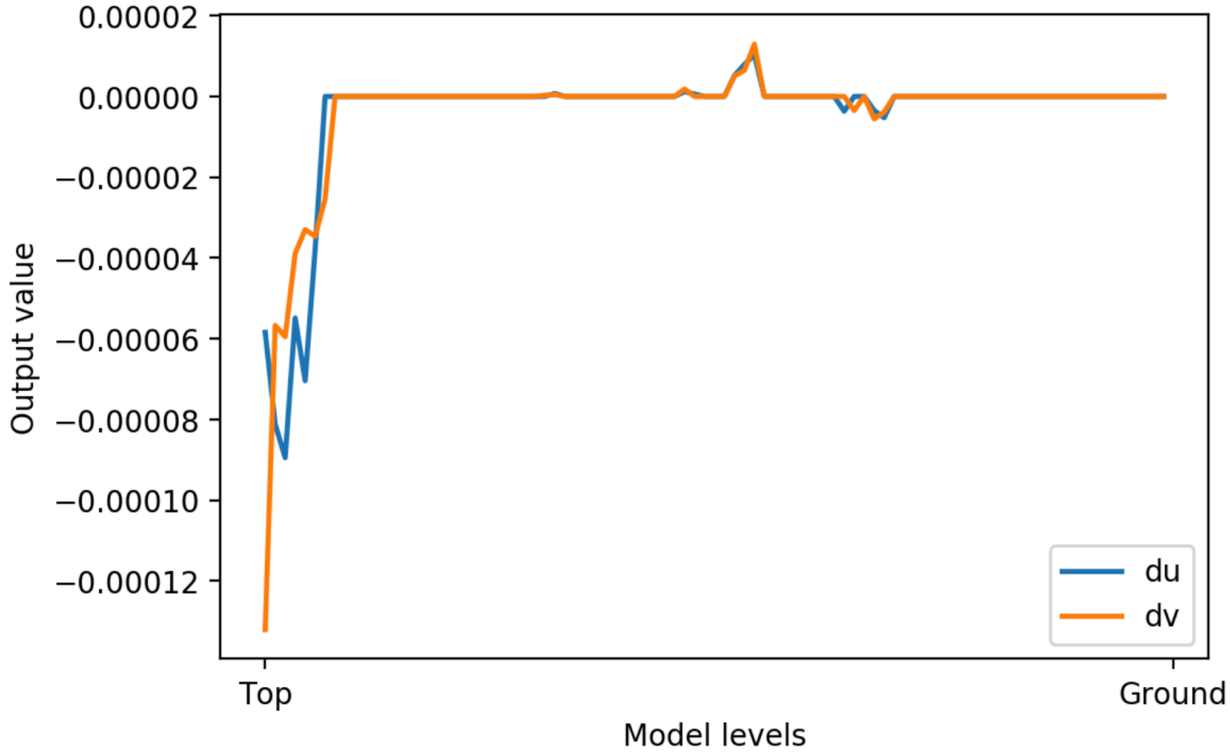
- Train with one year, test with another.

Data

Inputs



Outputs



How we normalise will affect both the ability to learn
and what features of the data to learn

Data: Less is more?

- Naive data:
  Input: u, v, T, pressure, half-level pressure, geopotential at all 91 vertical levels = 546 input variables
  Output: Increments du, dv = 182 outputs

- A bit of human learning/knowledge.

- Reduced data:
  Input: 3 variables x 63 levels + level data ~ 190
  Output: 2 variables x 63 levels = 126

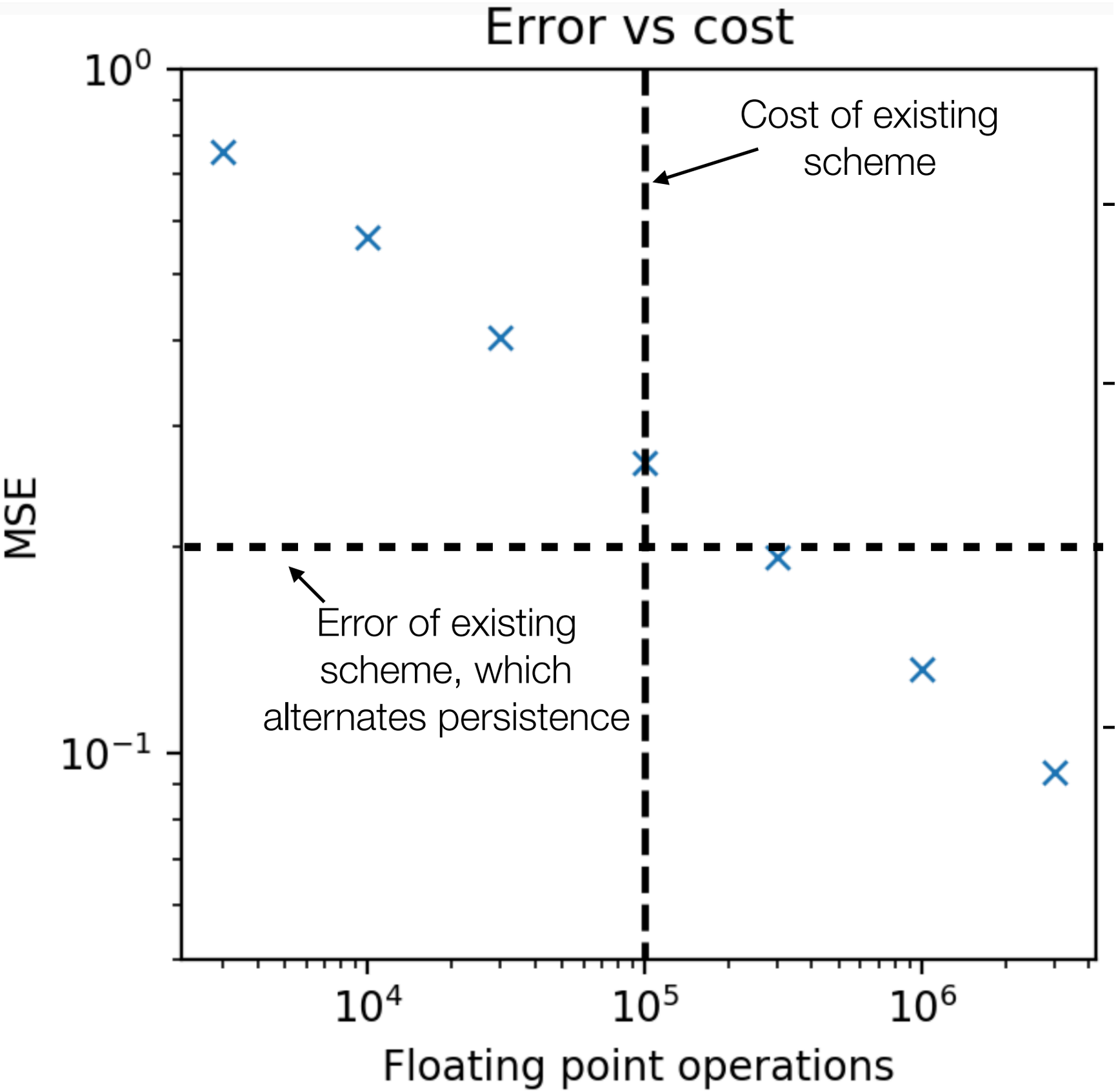
- Proof in the pudding. Converges to better model with fewer inputs.

NN design choices

- Focus on fixed-width, fully connected networks (more later).
    - Network has no a-priori knowledge that neighbours in vectors are neighbours in vertical level space.
- Search over width, depth, activation function, learning rate etc.
- Calculate FLOP cost and compare performance of equal cost NN.

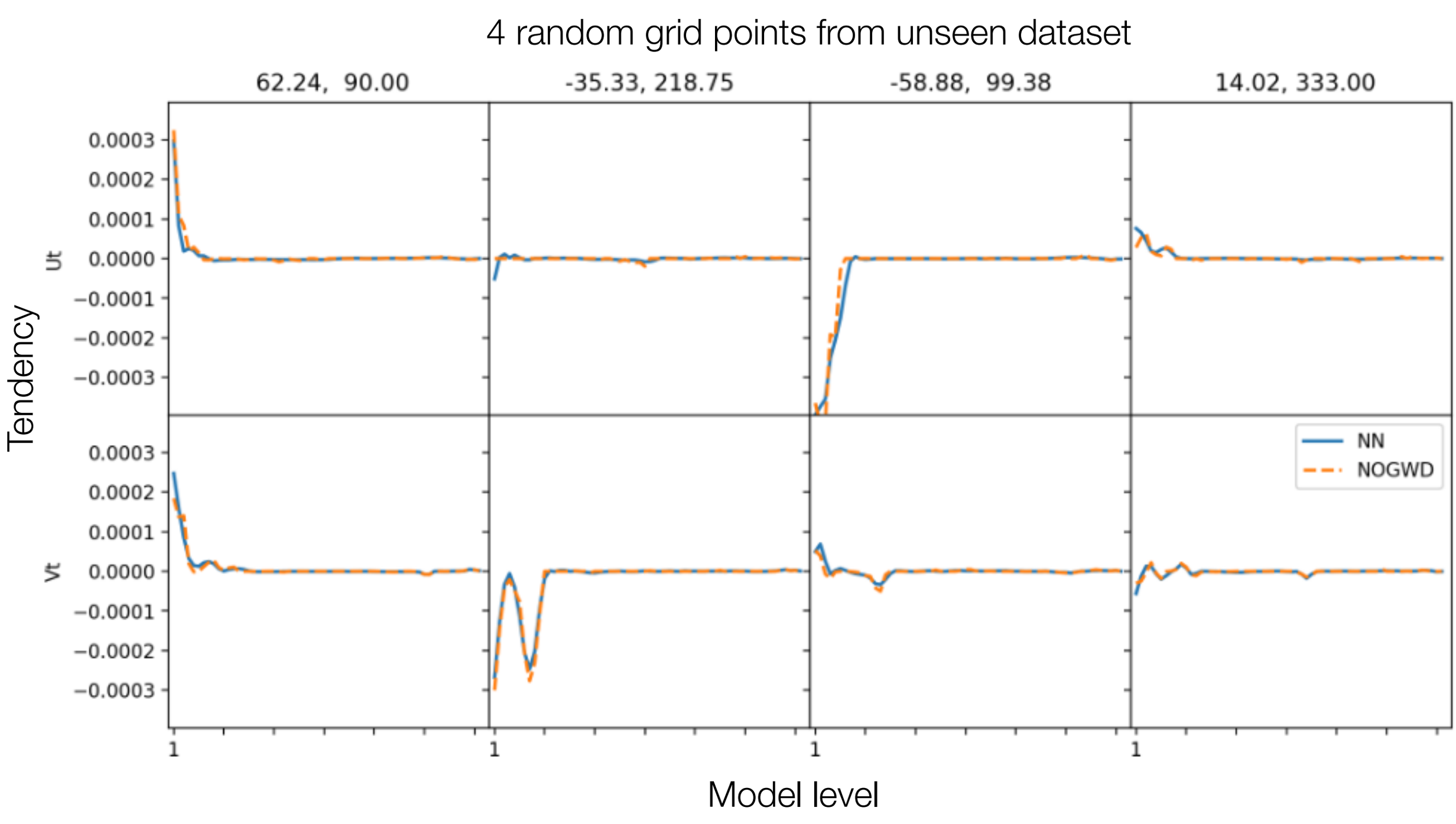
Physics constrained networks

- For the non-orographic gravity wave drag scheme, the tendency produced has no net momentum.
- i.e. $\int_{ground}^{top} udp = 0$
- In the current scheme this is achieved by dumping any net momentum on the highest vertical layer.
- For our networks we mimic this and train on (n-1) layers and put remaining momentum on top layer.
- This is similar to constrained architecture from Beucler et al.



- Better schemes = more expensive.
- No consistency in the optimal width/depth, varies with activation function and FLOP constraint.
- Optimal number of layers: 3-8

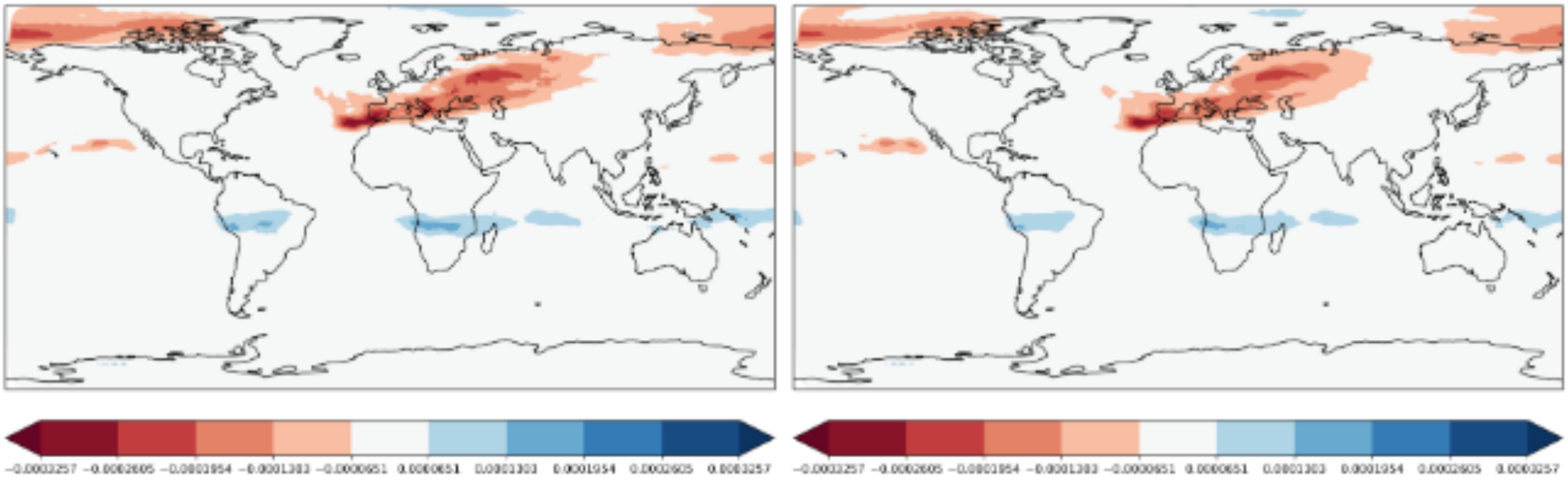
Results


4 random grid points from unseen dataset
62.24, 90.00
-35.33, 218.75
-58.88, 99.38
14.02, 333.00
Tendency
Ut
Vt
0.0003
0.0002
0.0001
0.0000
−0.0001
−0.0002
−0.0003
NN
NOGWD
1
Model level

Current scheme

10day averaged u-tendency @ 0.3hPa

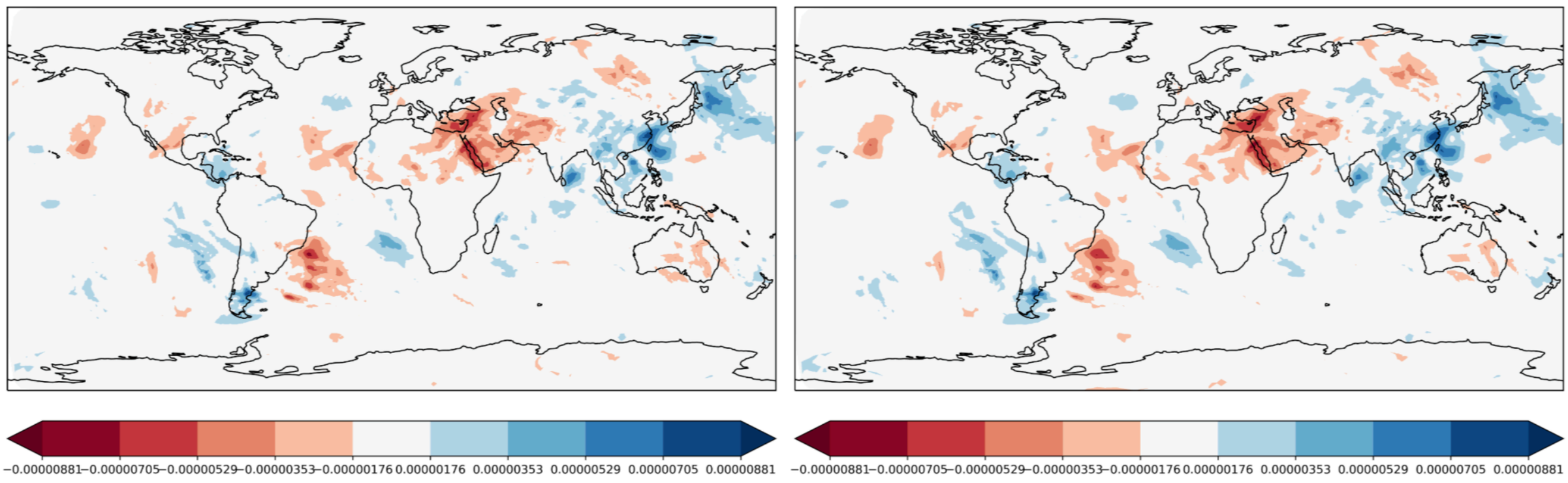
NN



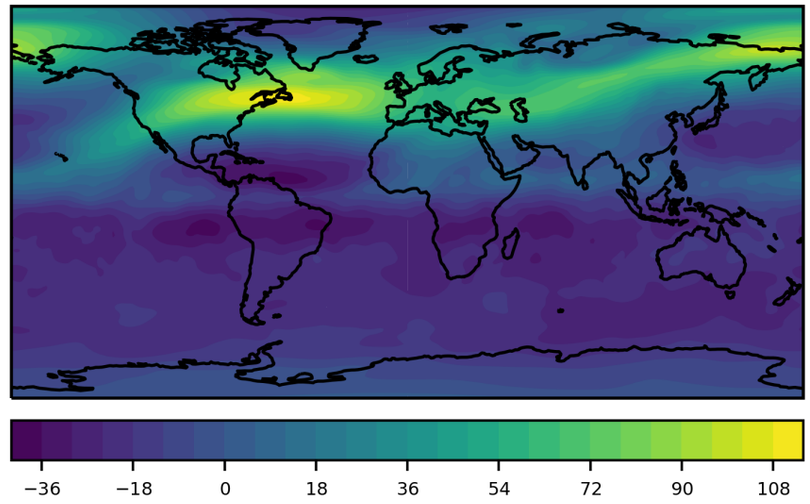
Current scheme

Instantaneous v-tendency @ 20hPa

NN



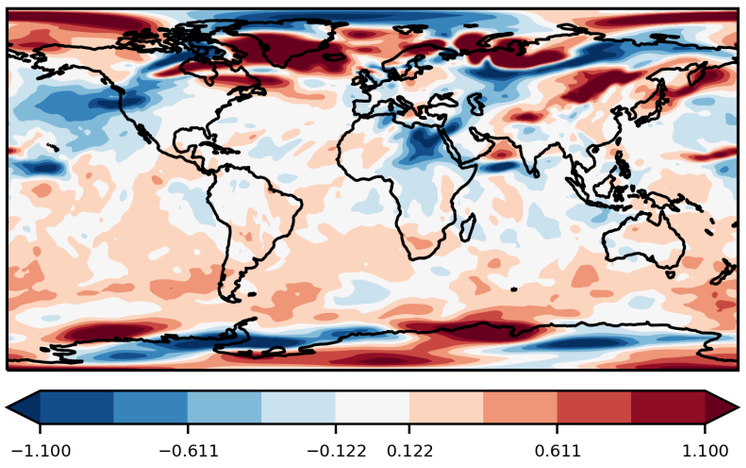
Coupled mode



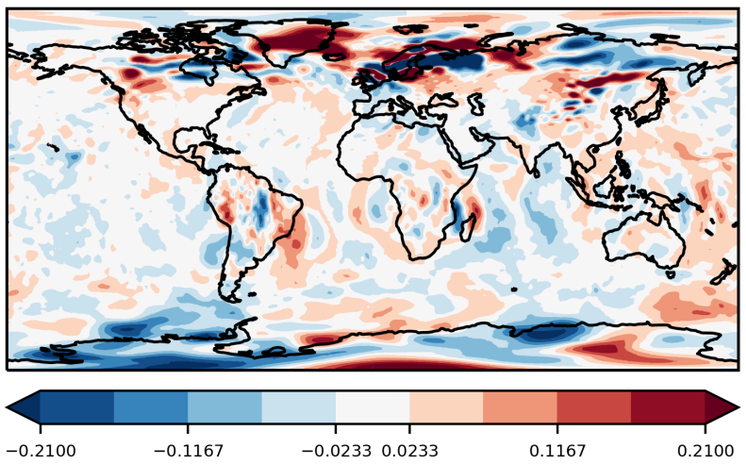
- Existing scheme replaced with NN inside IFS.

- Compare existing scheme run every timestep with:
  Scheme OFF, Scheme rerun every 2nd timestep (current setup), 3 NNs.

OFF, RMSE=0.434m/s

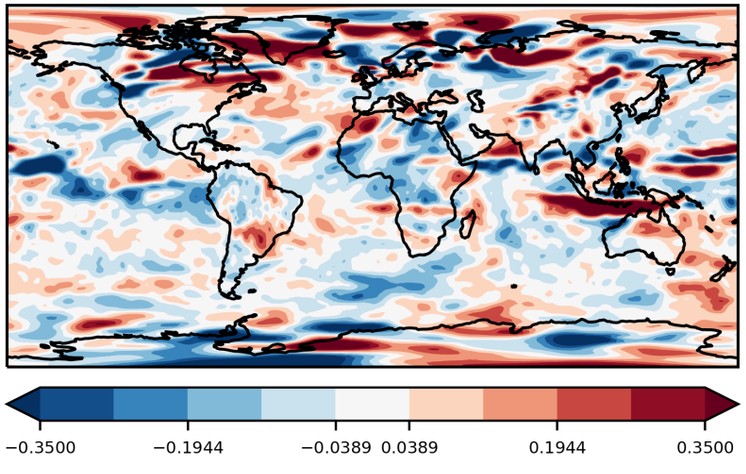


Default, RMSE=0.087m/s

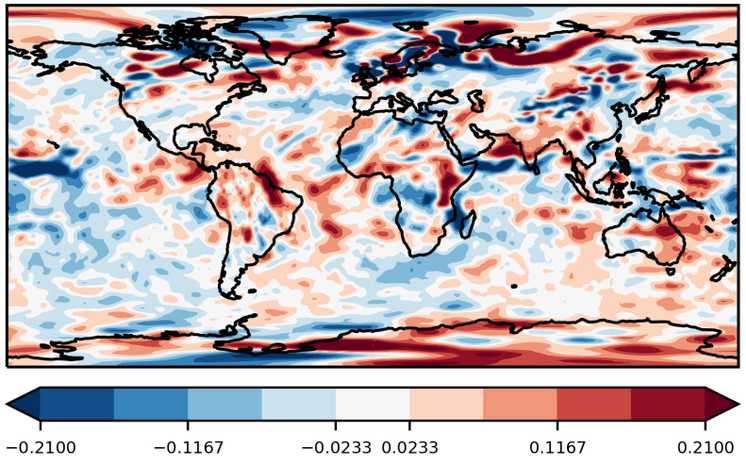


24hr forecast
U winds @5hP
~36km

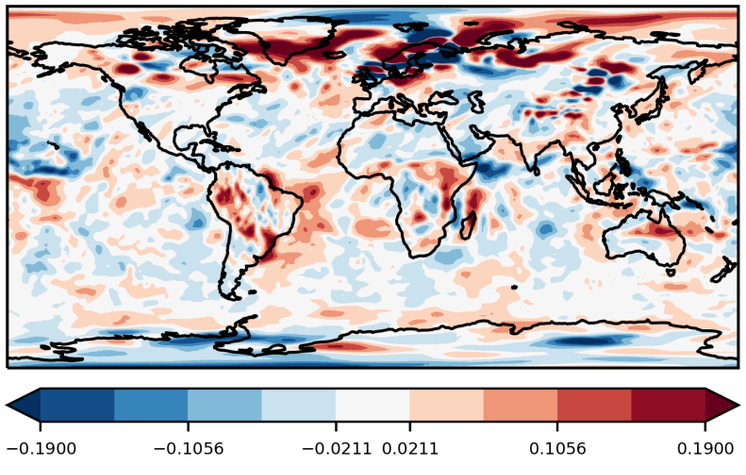
30k flops, RMSE=0.156m/s



300k flops, RMSE=0.095m/s



3m flops, RMSE=0.078m/s



Marginal difference between schemes of dramatically varying cost.

What about longer integrations?

- Run for year long simulations each starting from 1-Nov, for 6 years.

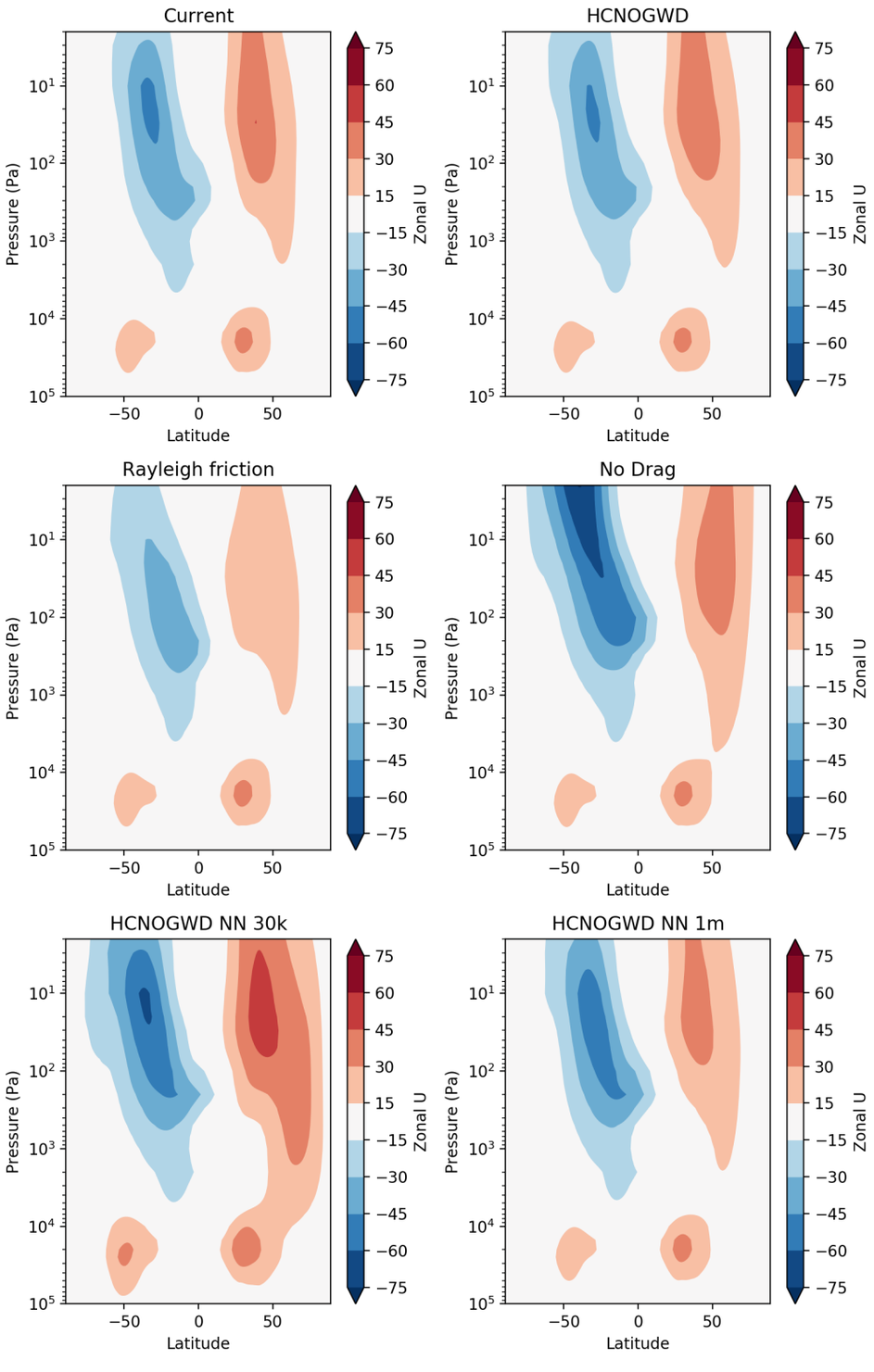
- First models crash!

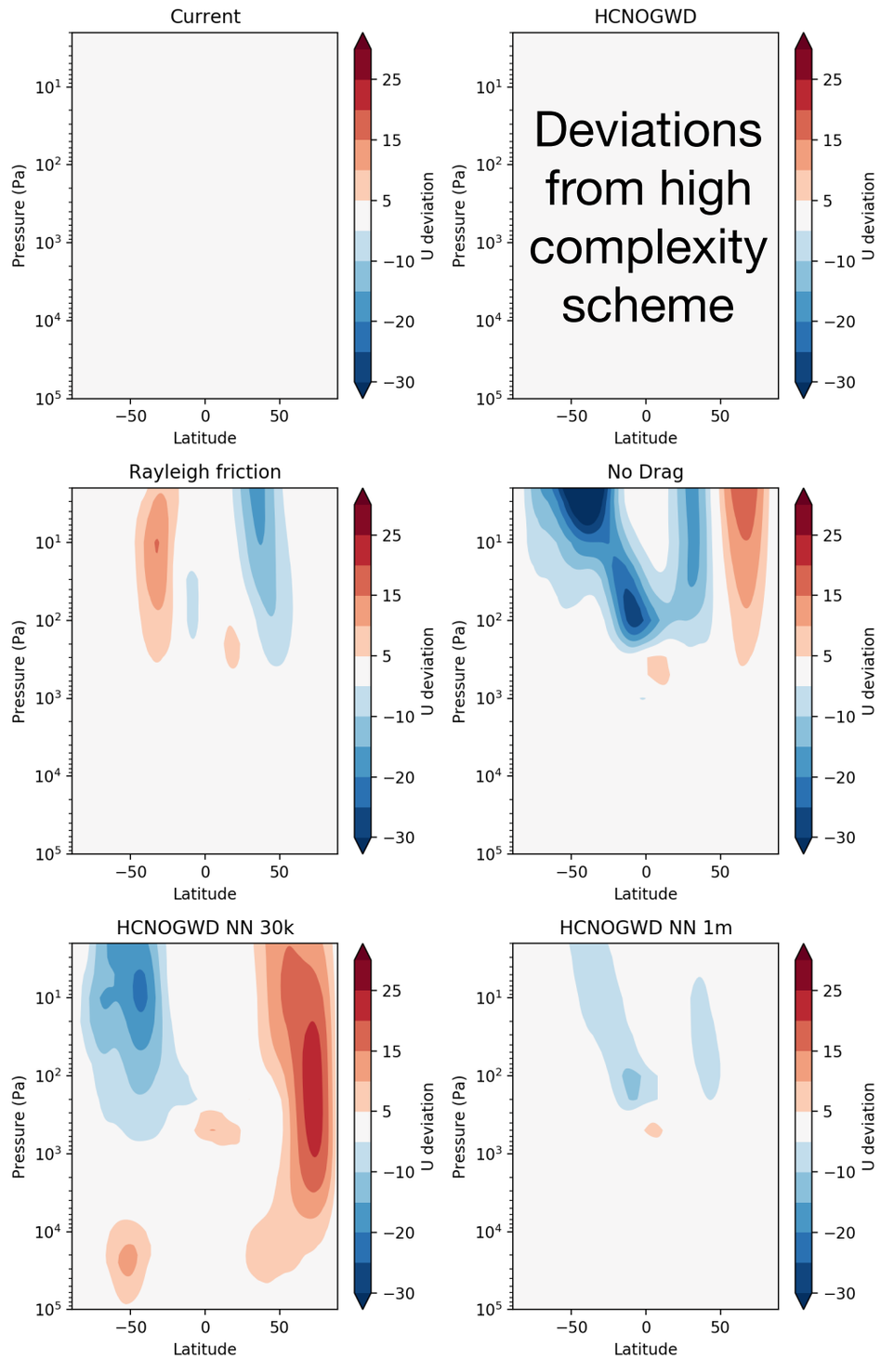
- Our physics-constrained network is unstable, the highest layers are densely packed in pressure, so small errors in lower atmosphere result in large tendencies at the top layer.

- Long-term: retrain including top layer in loss (i.e. Beucler et al.)
  For now: predict entire column, small errors in momentum conservation.

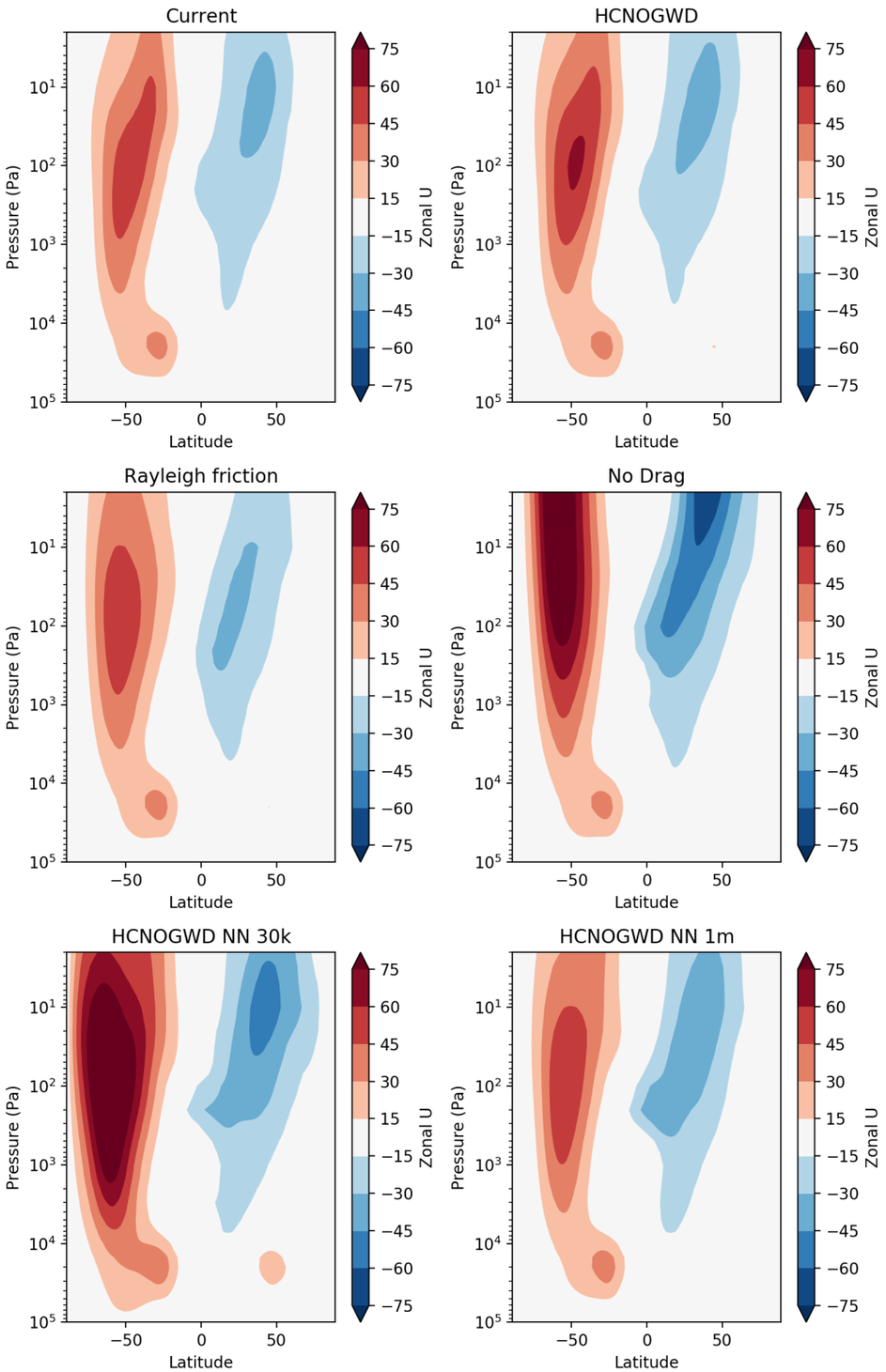
- Examine zonal structure of the atmosphere.

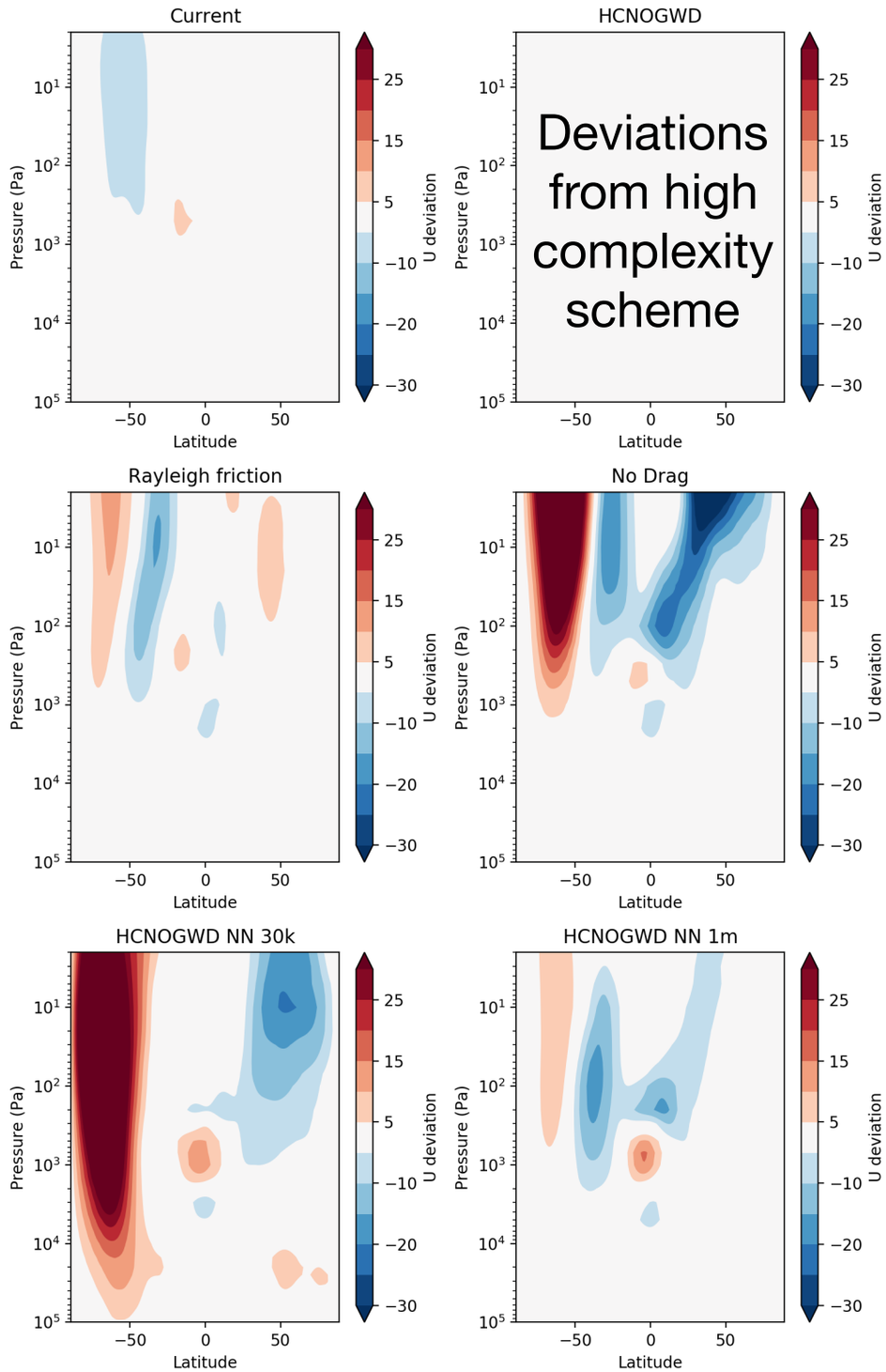
DJF Zonal velocities



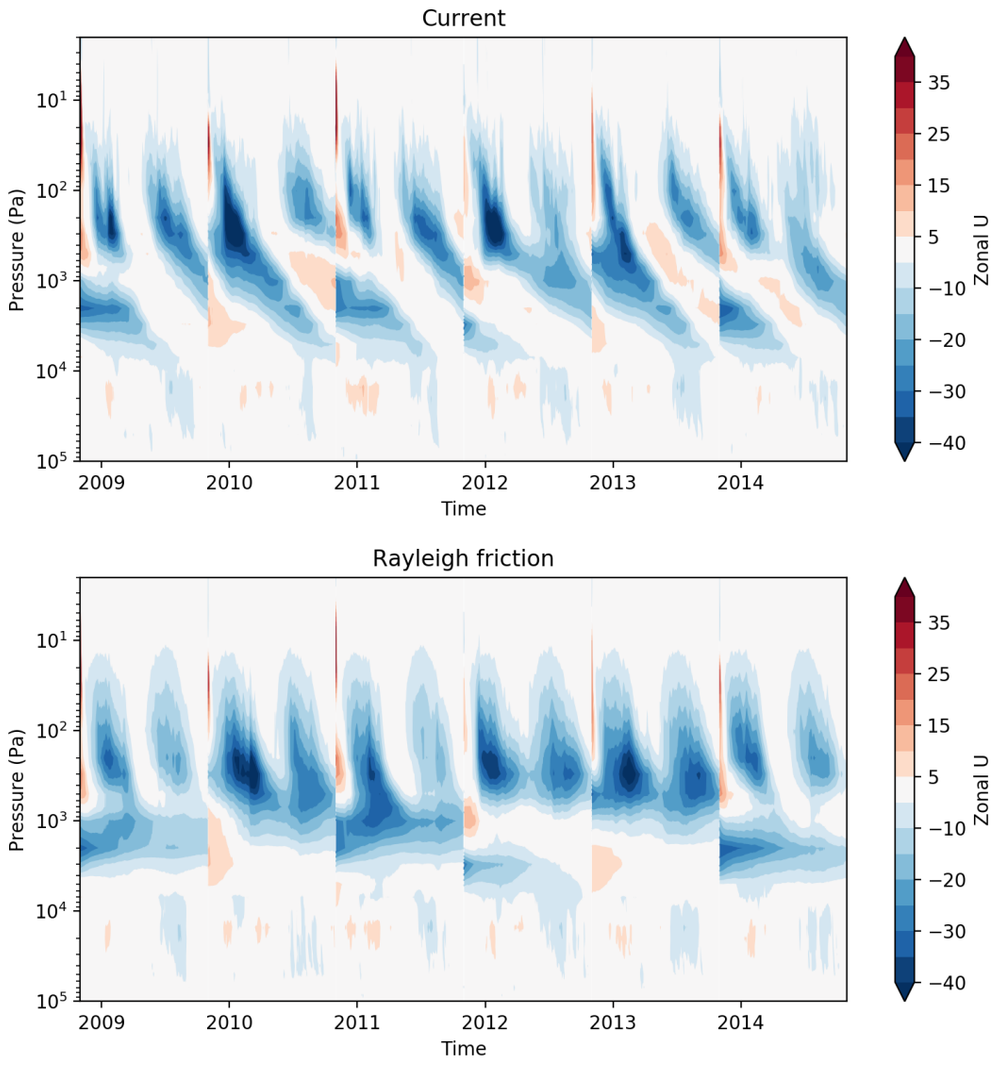


JJA Zonal velocities

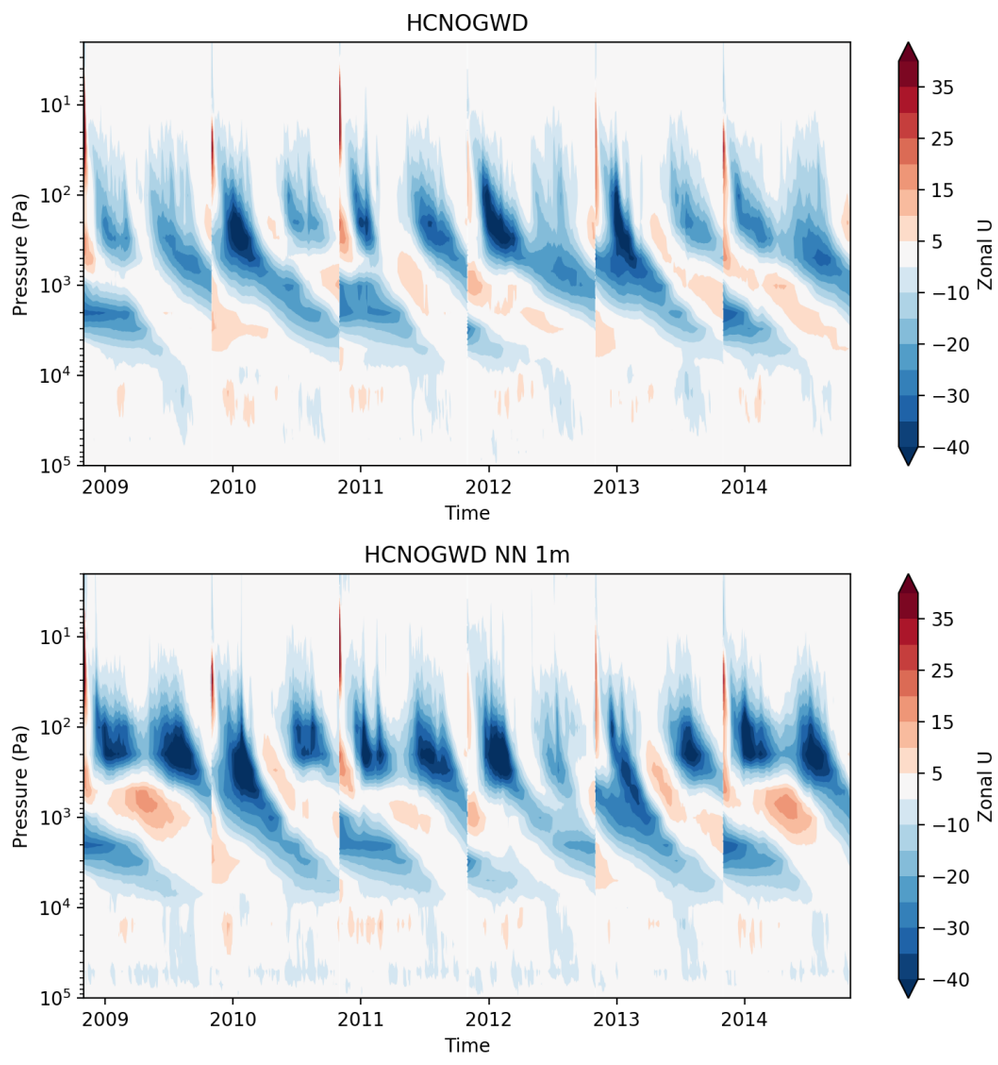




Quasi-biennial oscillation



Rayleigh friction



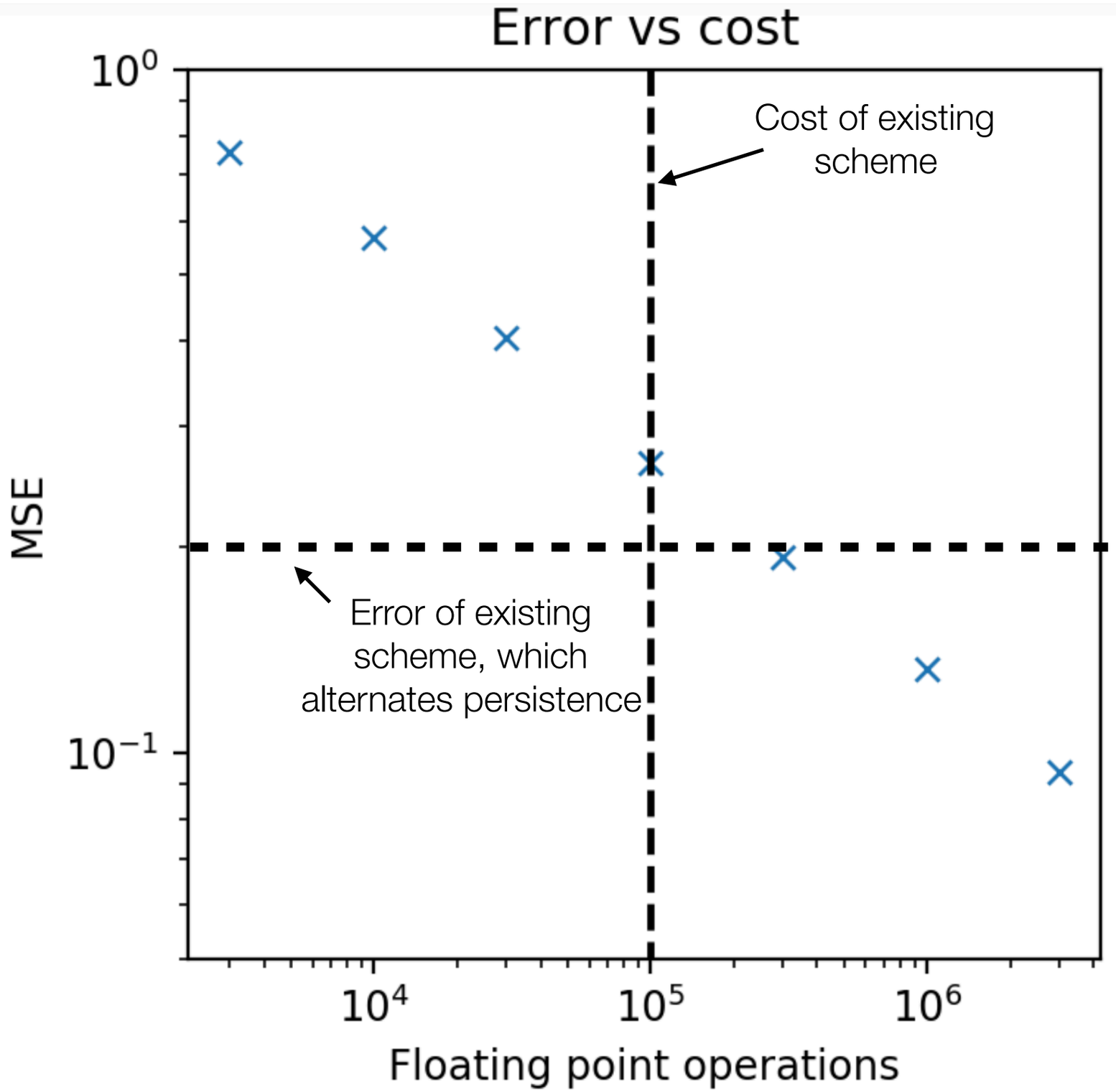
HCNOGWD NN 1m

Jet averaged over [-10,10]

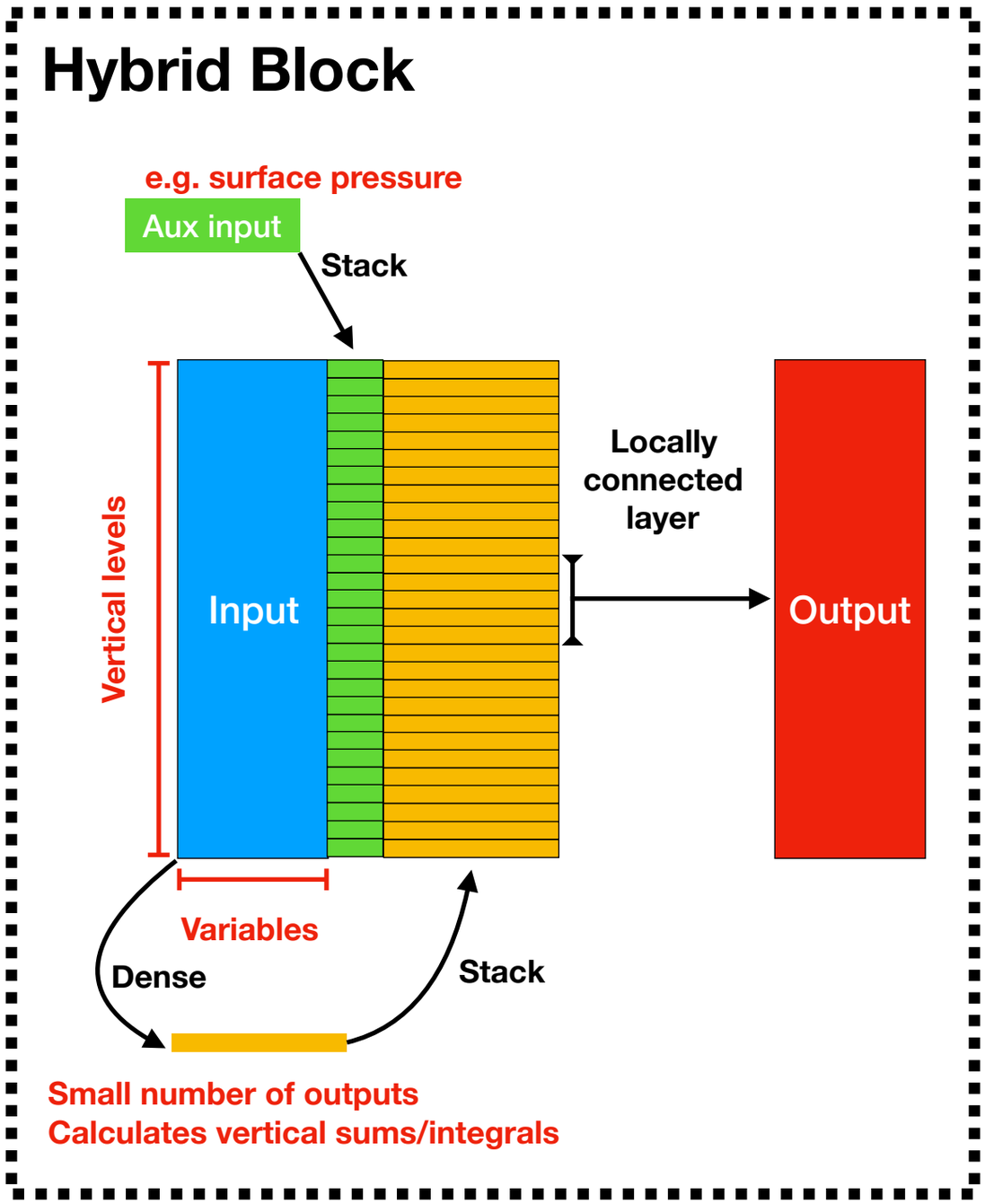
Jet slightly too strong in NN but captures decent (unlike Rayleigh friction)

but Matthew…

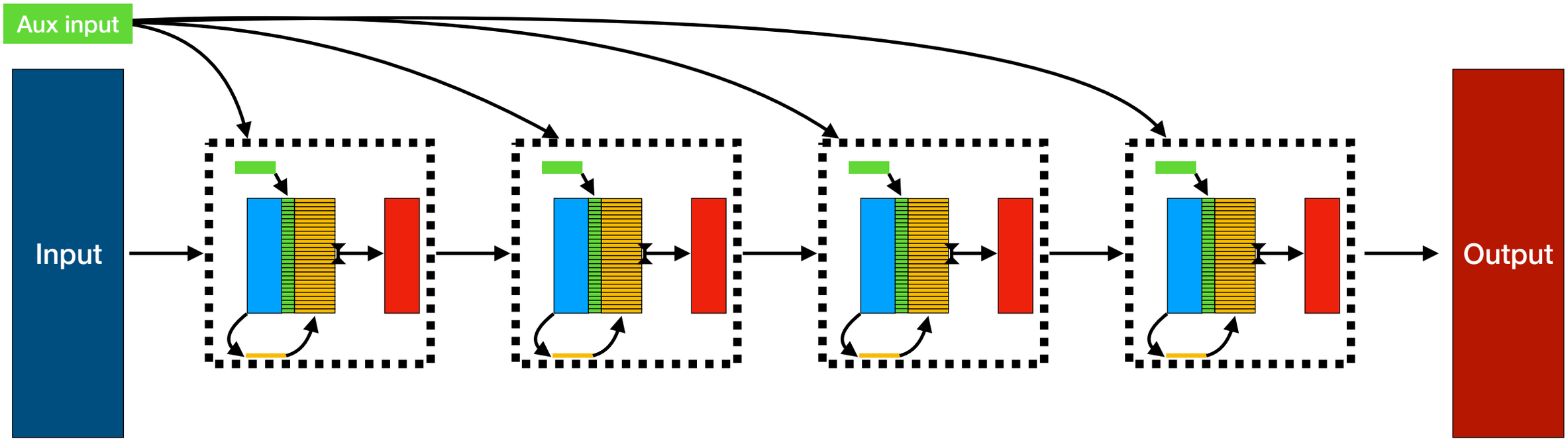
What about the cost?



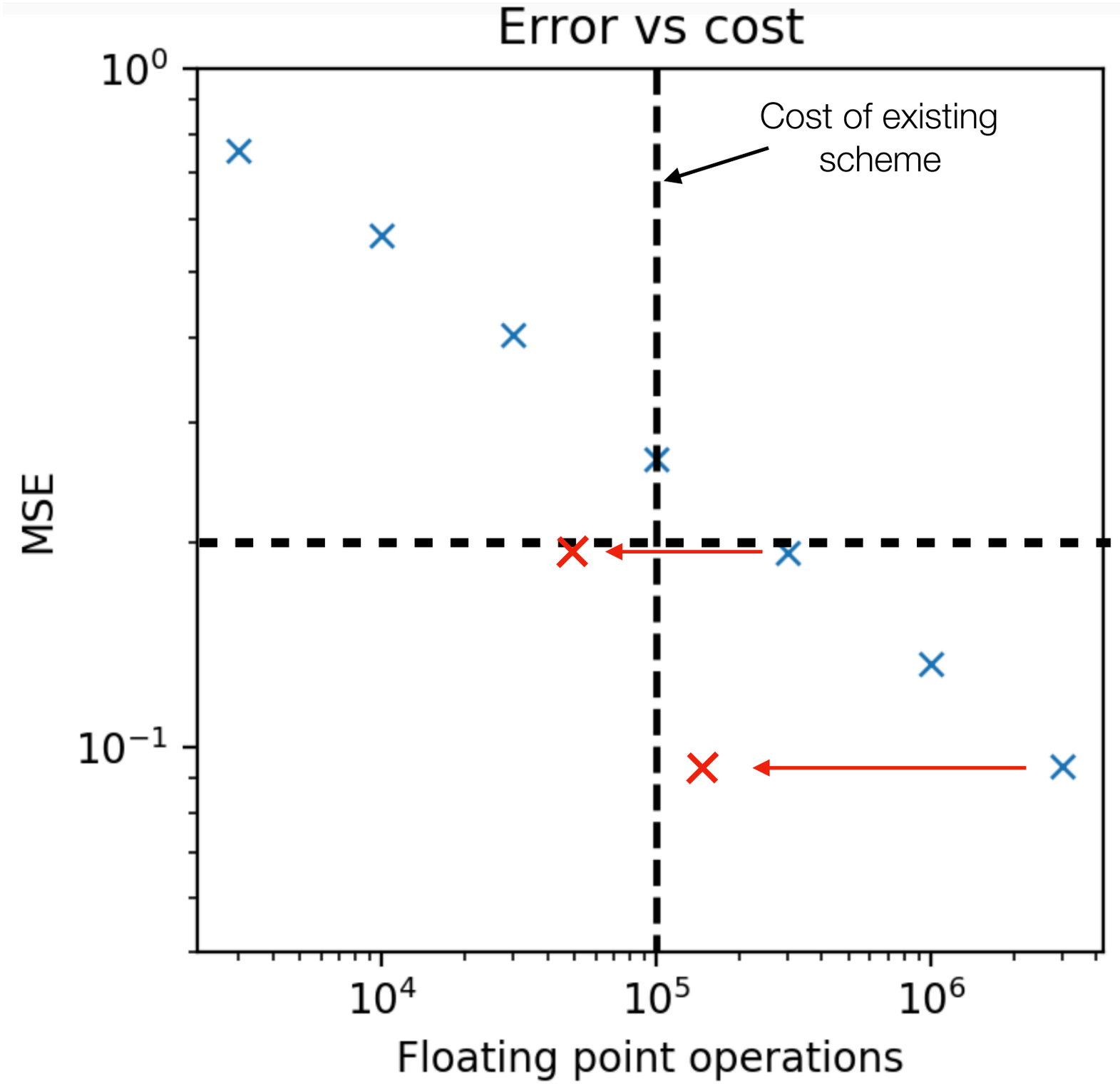
What about the cost?



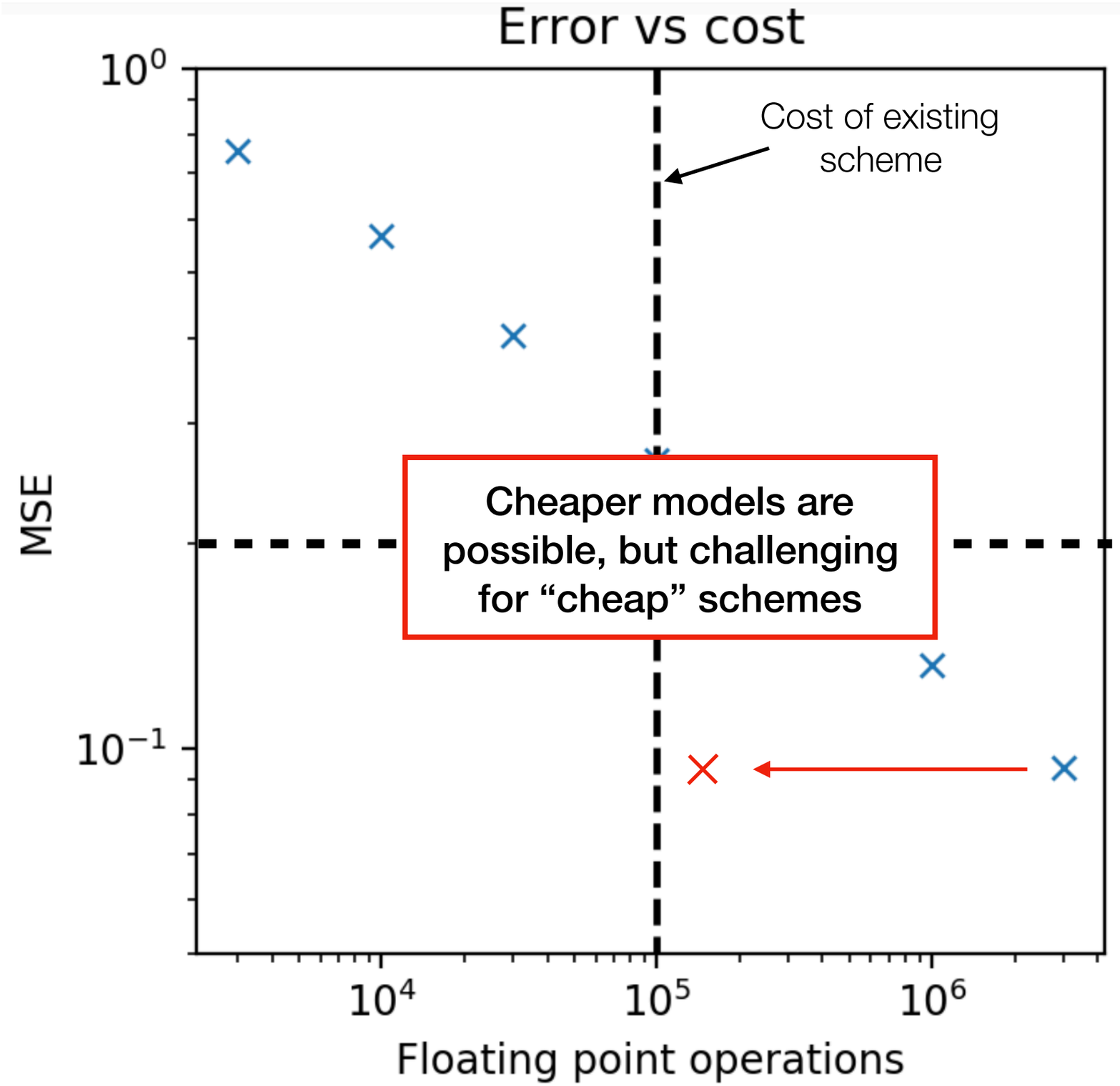
What about the cost?



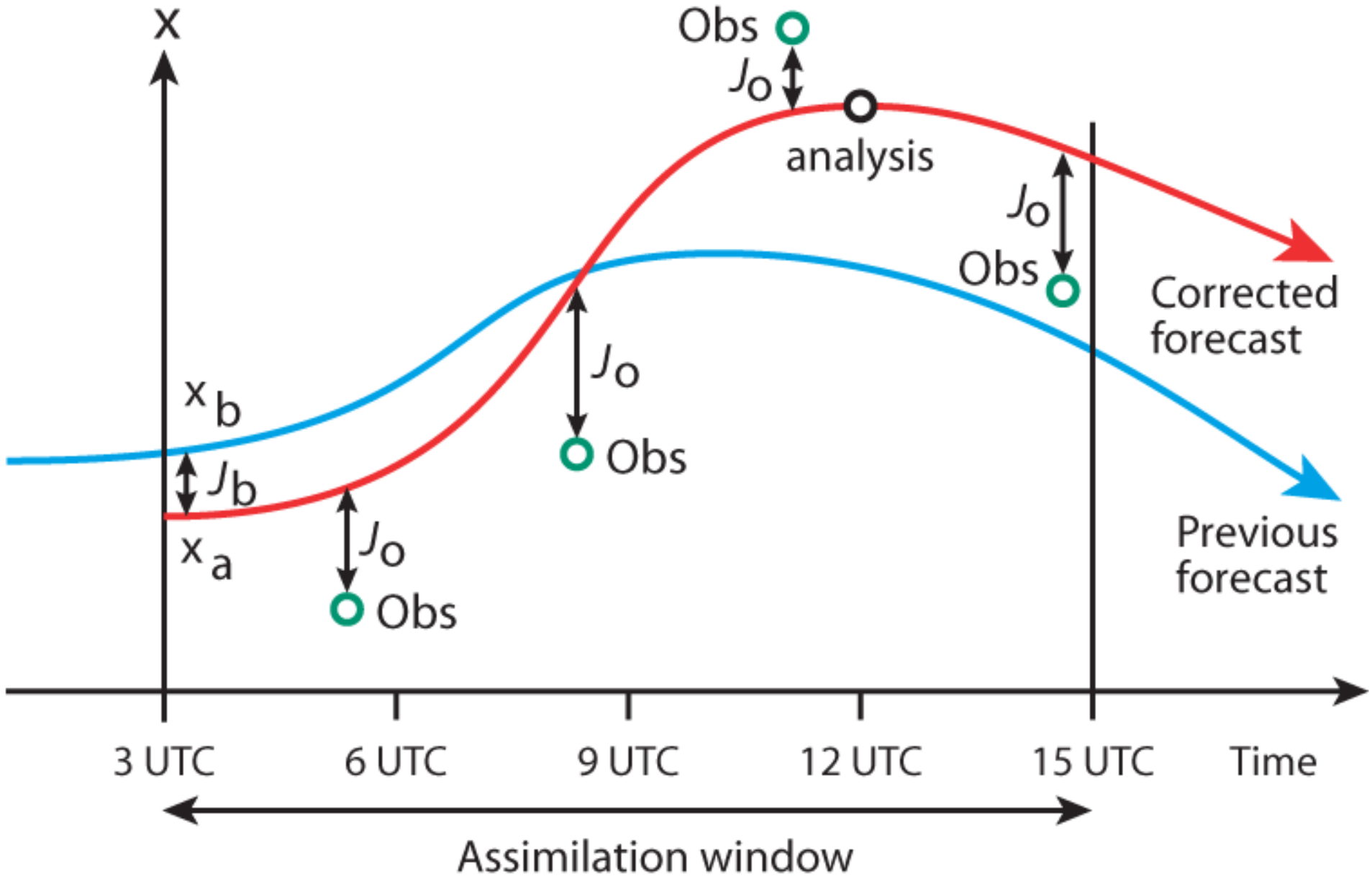
What about the cost?



What about the cost?

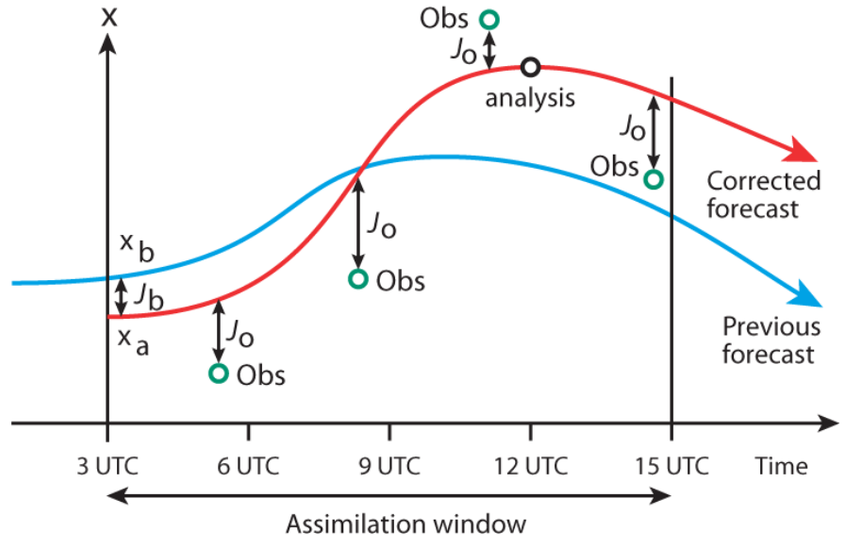


Data assimilation



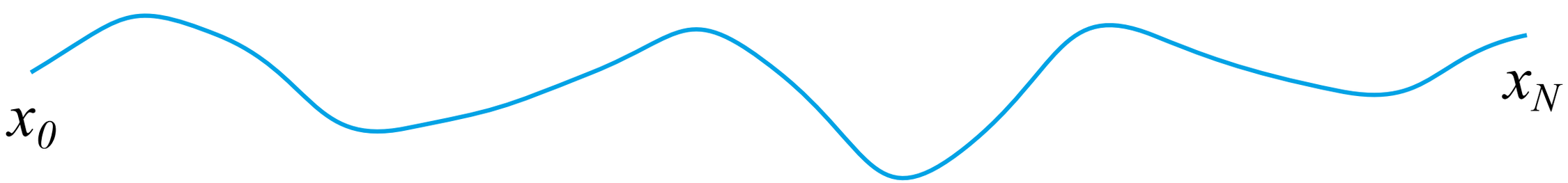
Credit: ECMWF

4D-var approach

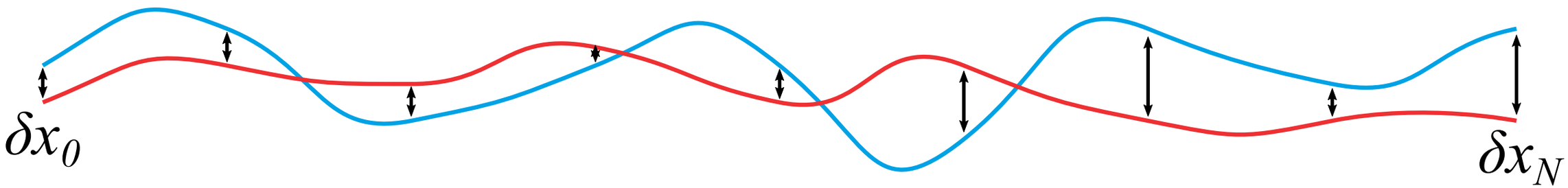


assimilation window

nonlinear model

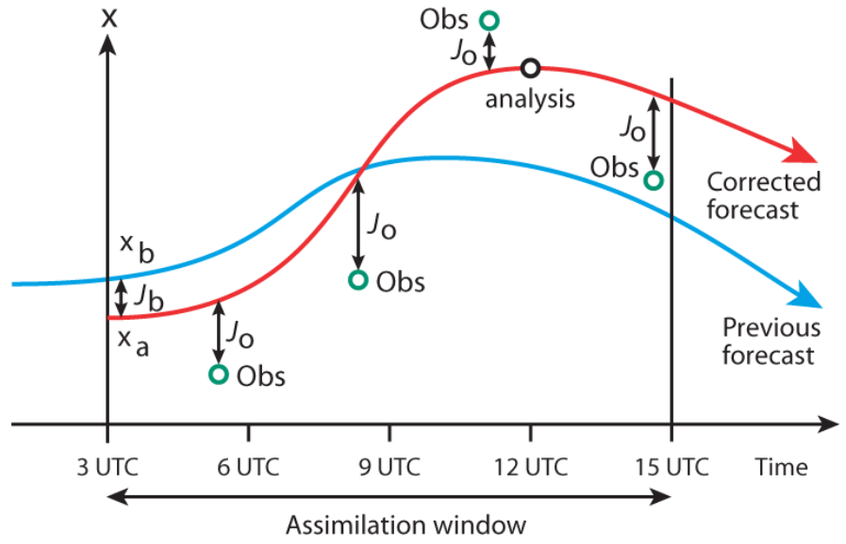


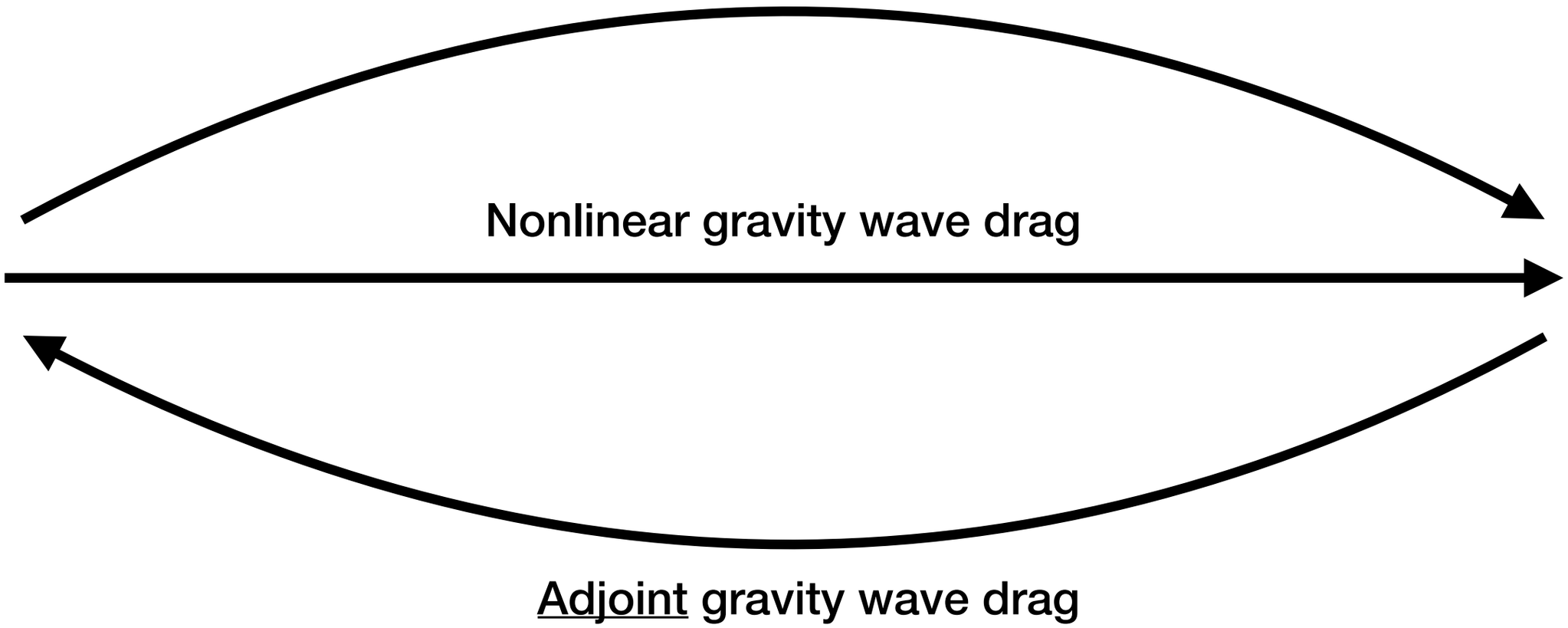
tangent-linear model



adjoint model

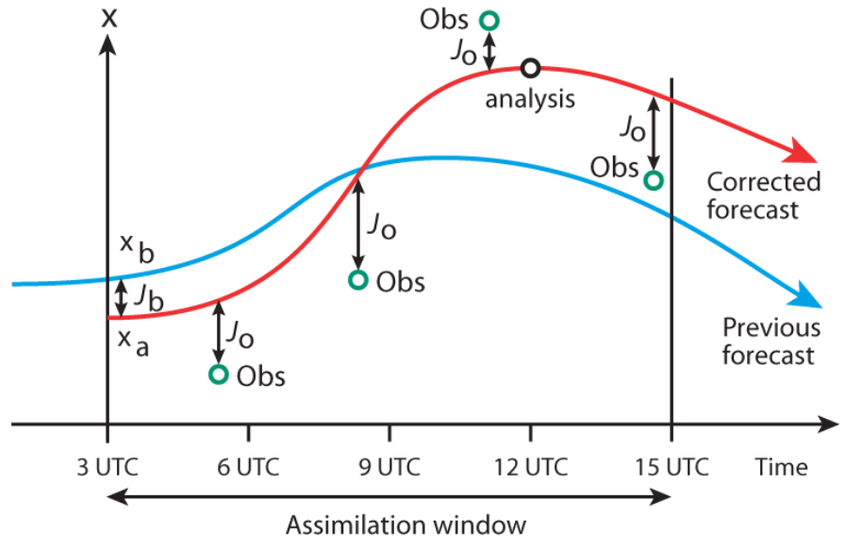
What is needed?





These are hard to develop & maintain. Almost always involve simplifications.

Our experiments



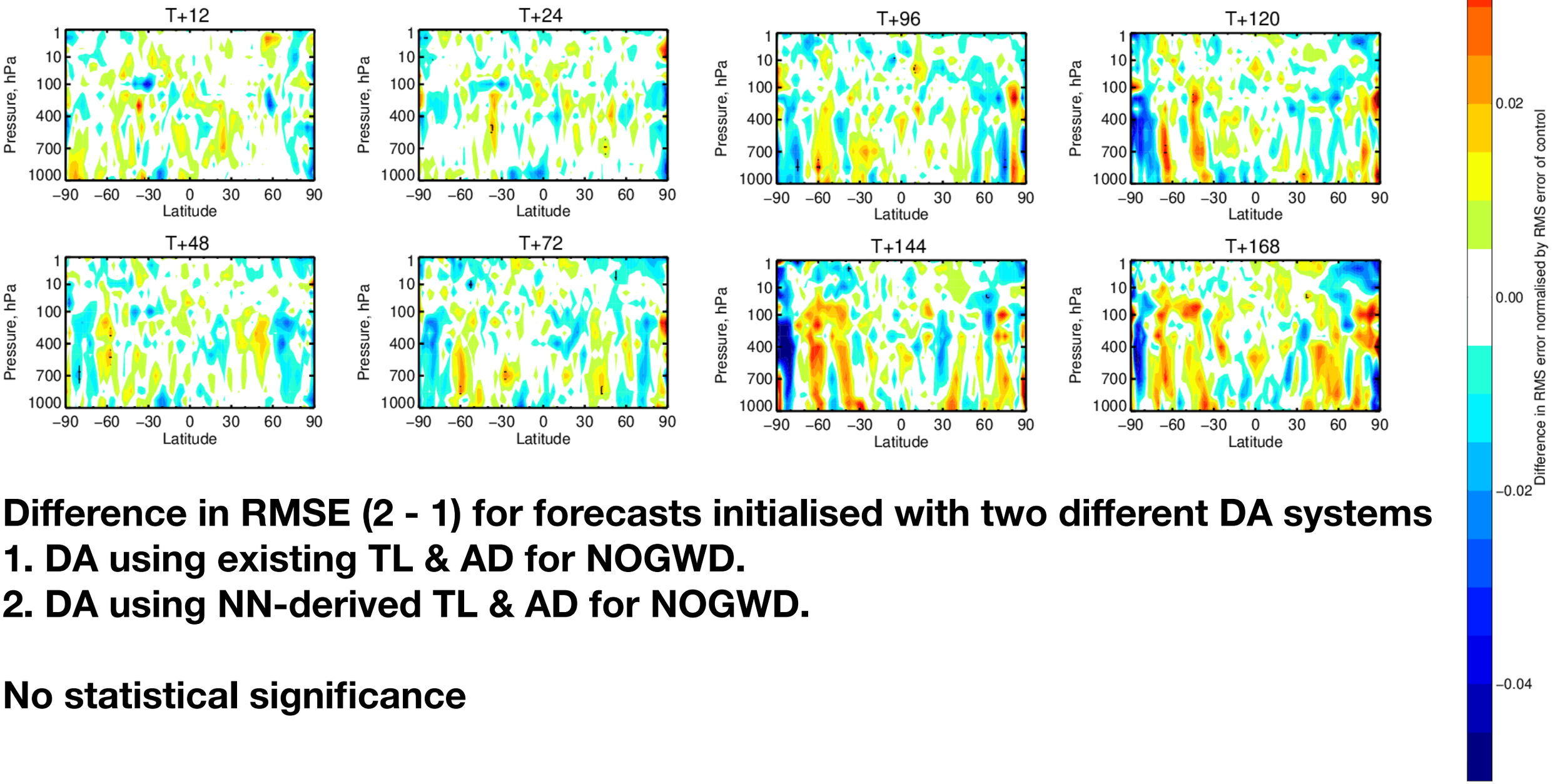
NN Tangent linear gravity wave drag

Nonlinear gravity wave drag

NN Adjoint gravity wave drag

Use existing nonlinear gravity wave drag. Use the NN to calculate TL & AD.

Our experiments



Difference in RMSE (2 - 1) for forecasts initialised with two different DA systems

1. DA using existing TL & AD for NOGWD.

2. DA using NN-derived TL & AD for NOGWD.

No statistical significance

NOGWD Summary

- Can we emulate parameterisation schemes? Yes

- Are they cheaper than the originals? Maybe

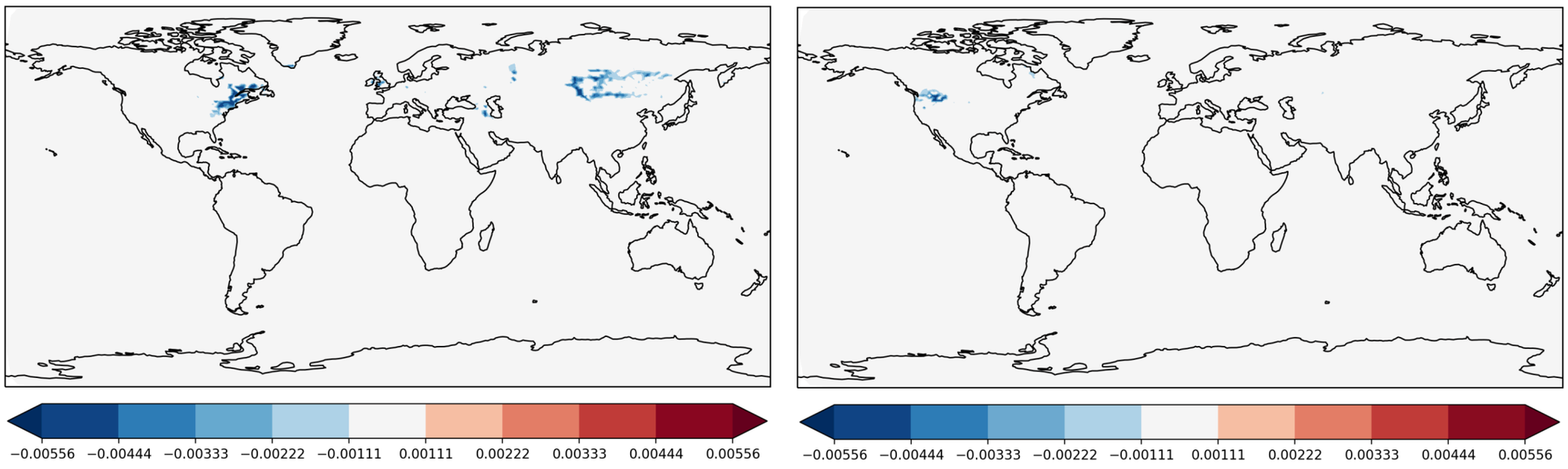
- Can this help with data assimilation? Yes

Orographic Gravity wave drag

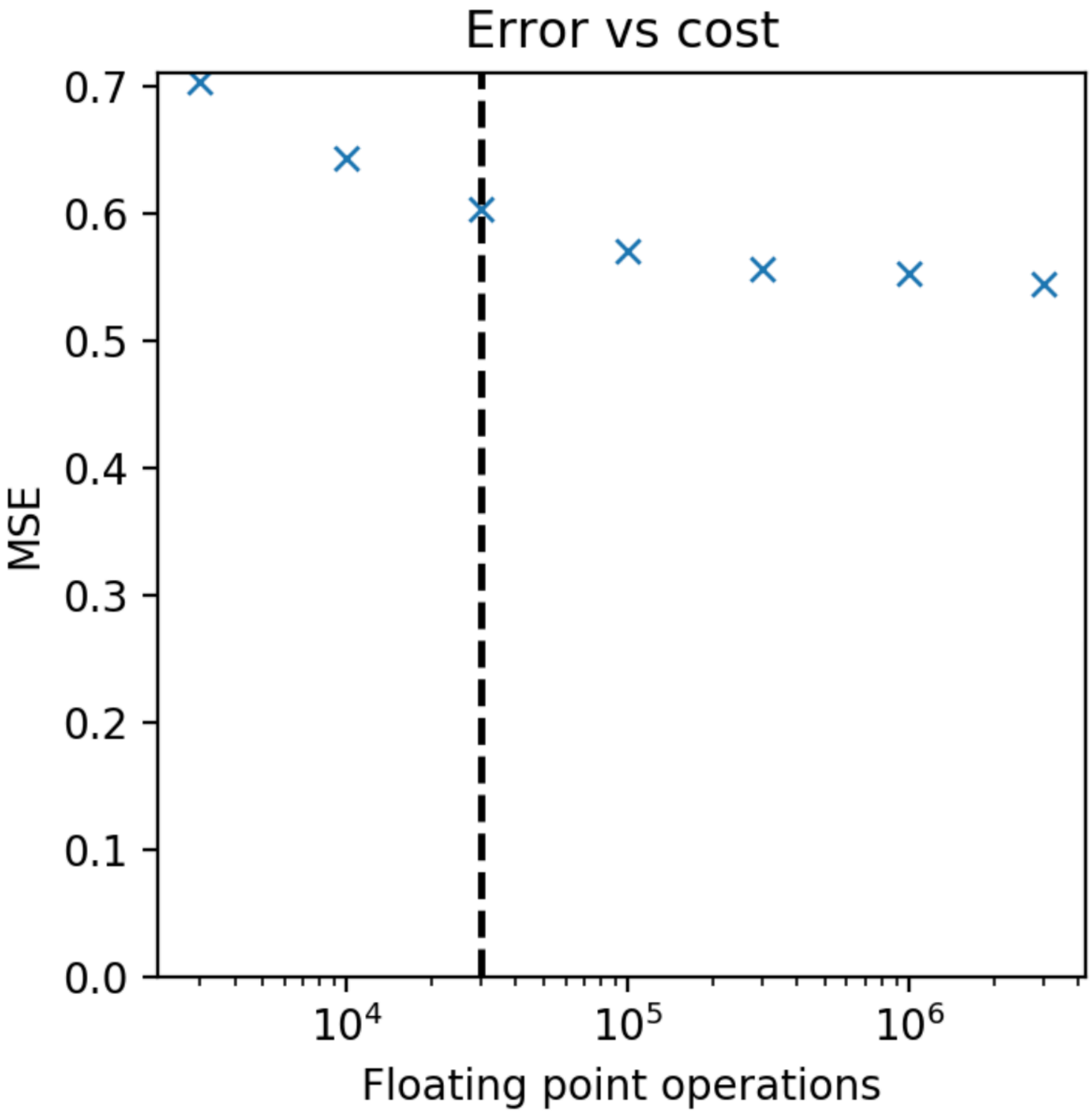
- Capture the impact of orographic features which are smaller than the grid scale.

- Broadly very similar scheme from a NN perspective, similar input and output vectors.

- Key difference: strong localisation in outputs.

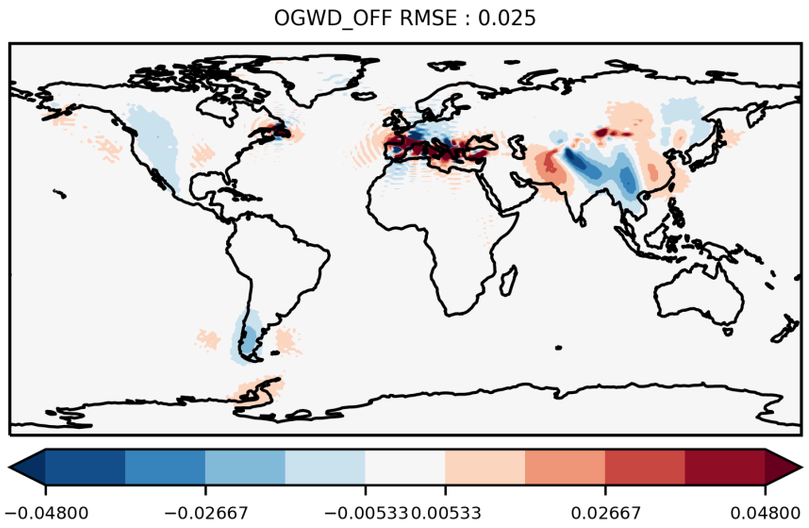


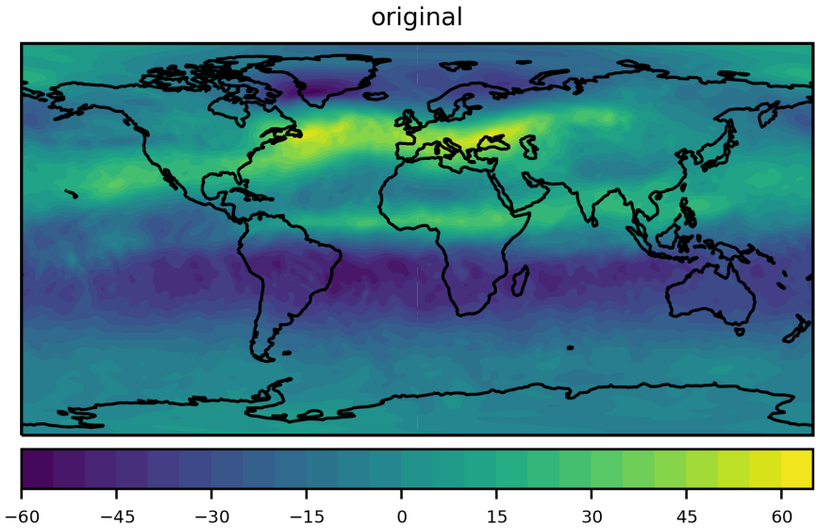
Different time-steps (dynamic conditions) produce very different sparse outputs.



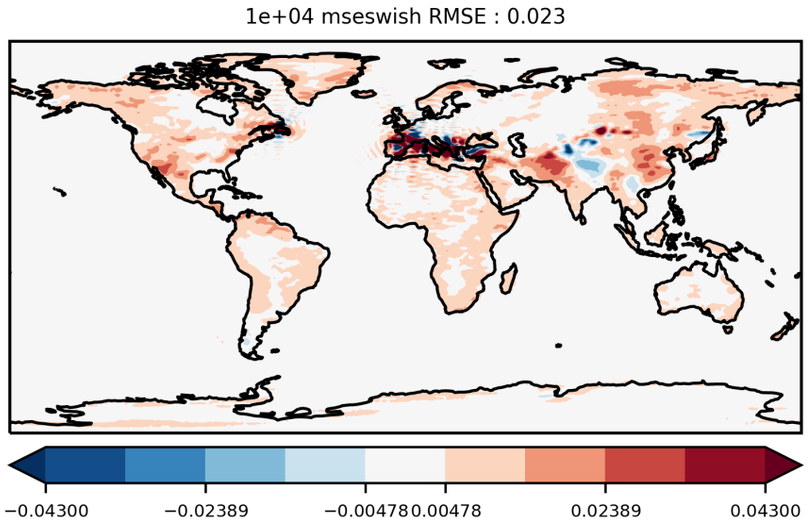
- Error stagnates.
- Compare with a persistence error of O(1).
- Error seems large, try coupled?

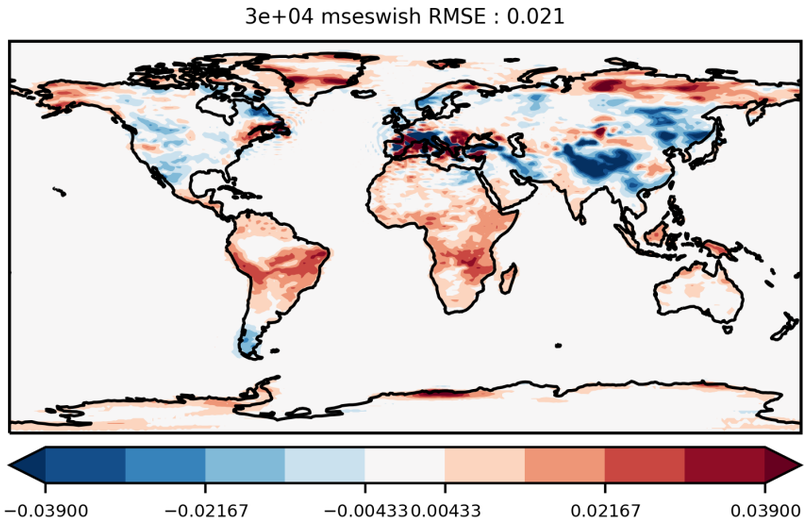
Coupled mode, T = 1

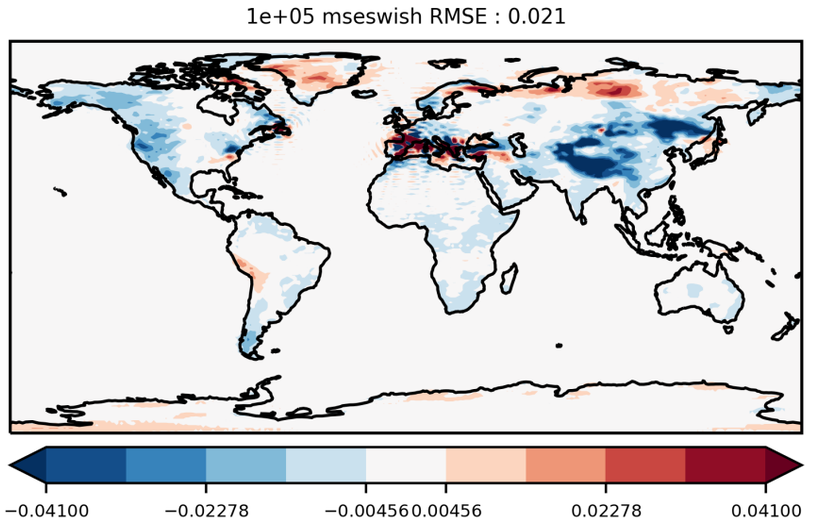




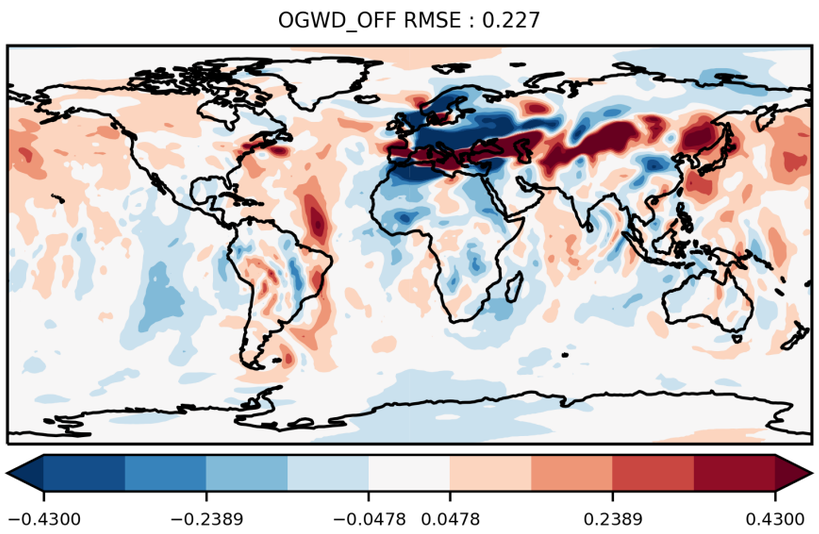
U velocity
5hP

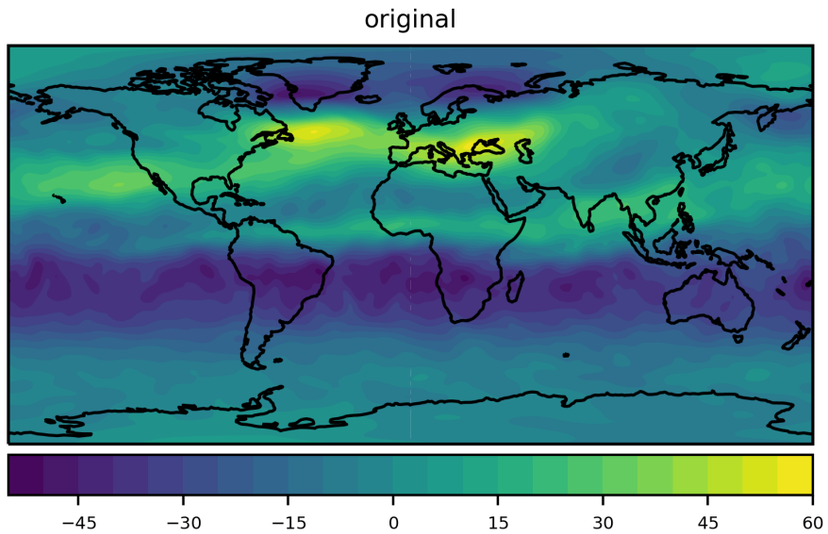




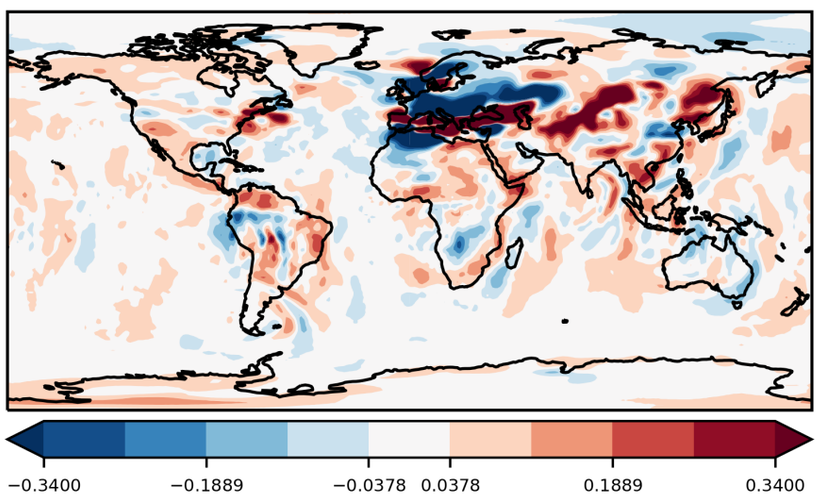


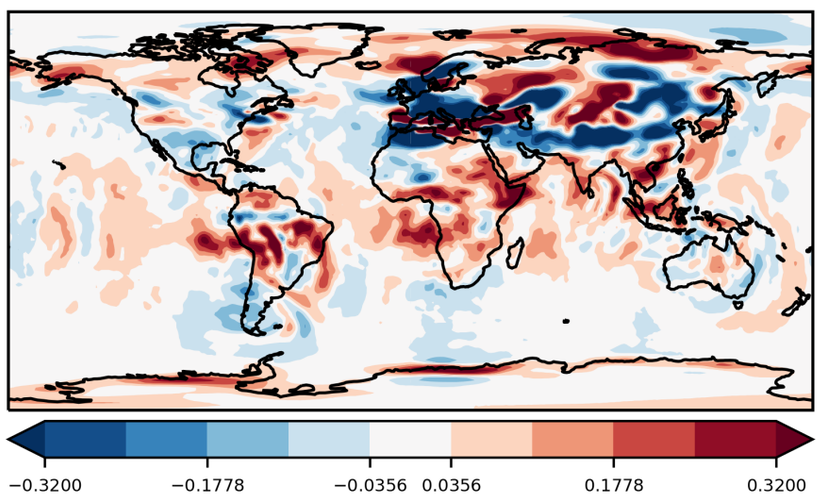
Coupled mode, T = 24

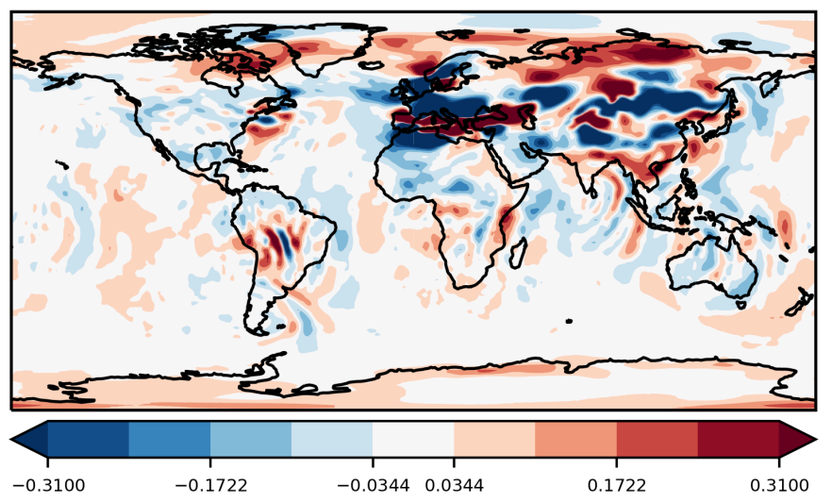




U velocity
5hP

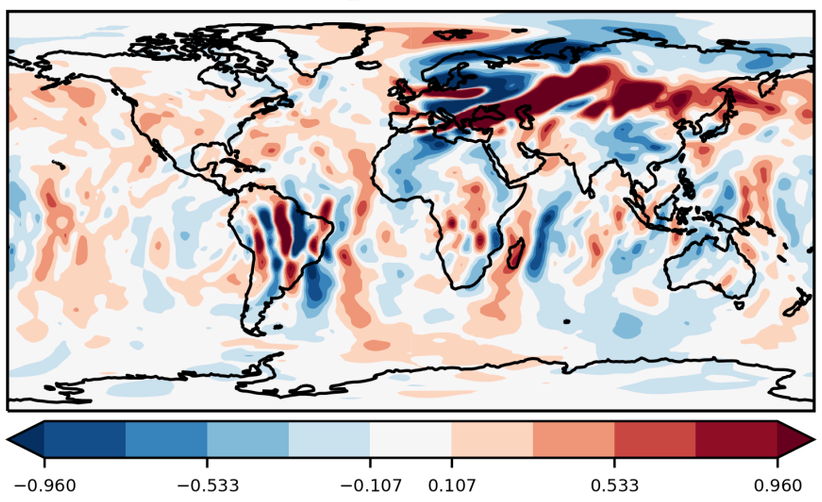




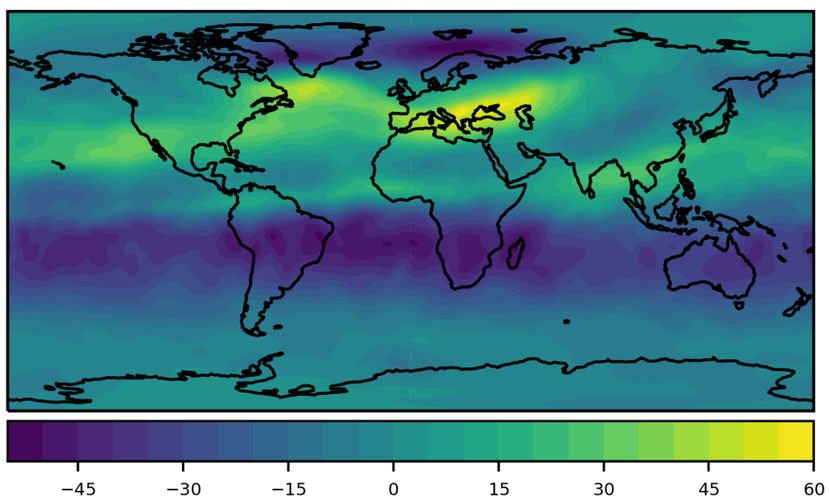


Coupled mode, T = 48

OGWD_OFF RMSE : 0.497

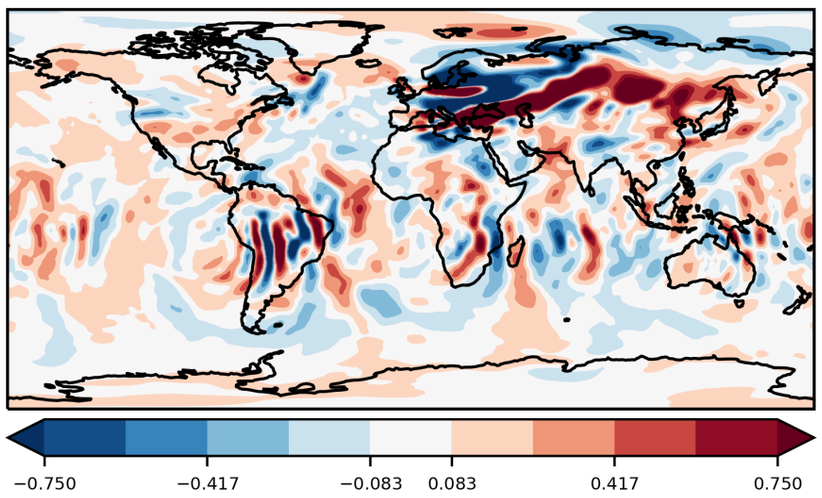


original

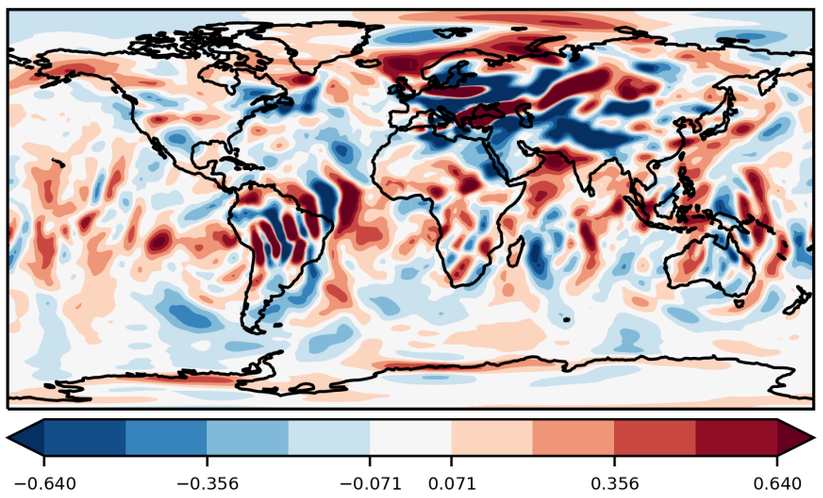


U velocity
5hP

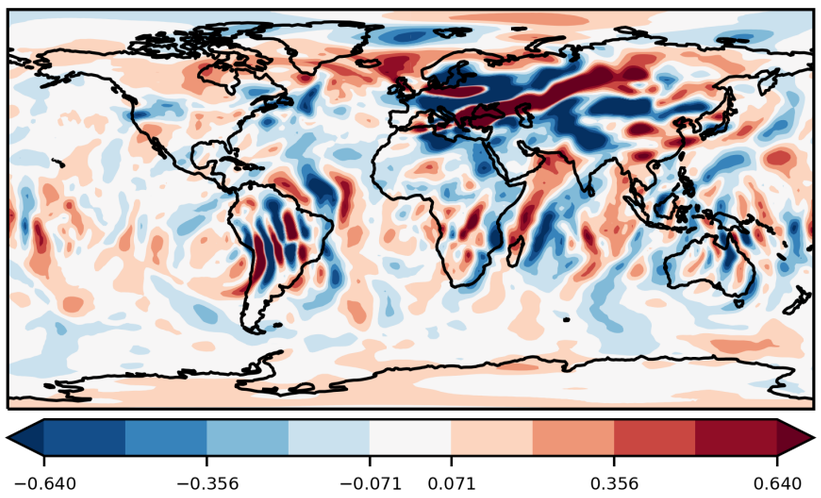
1e+04 mseswish RMSE : 0.393



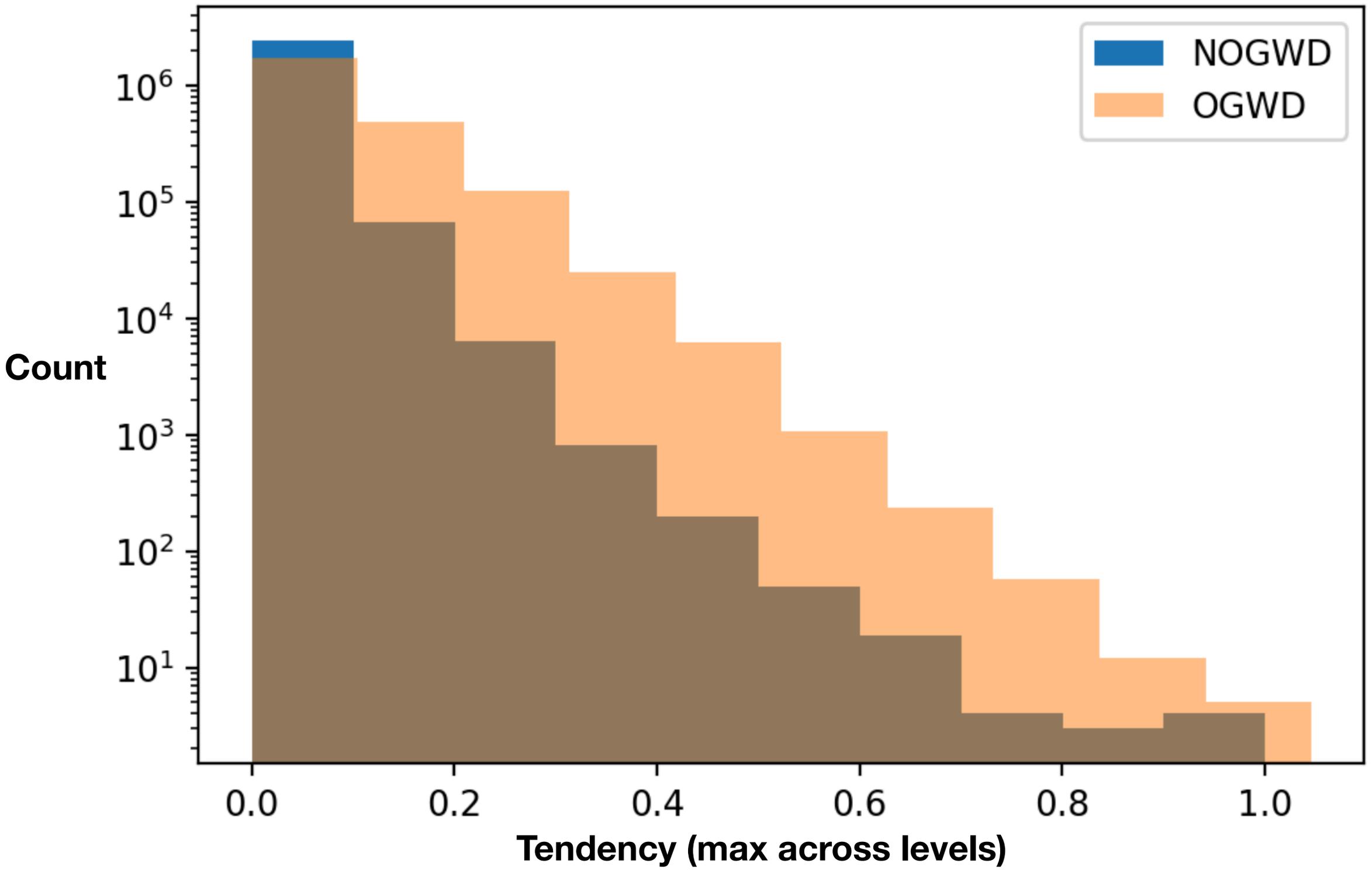
3e+04 mseswish RMSE : 0.342



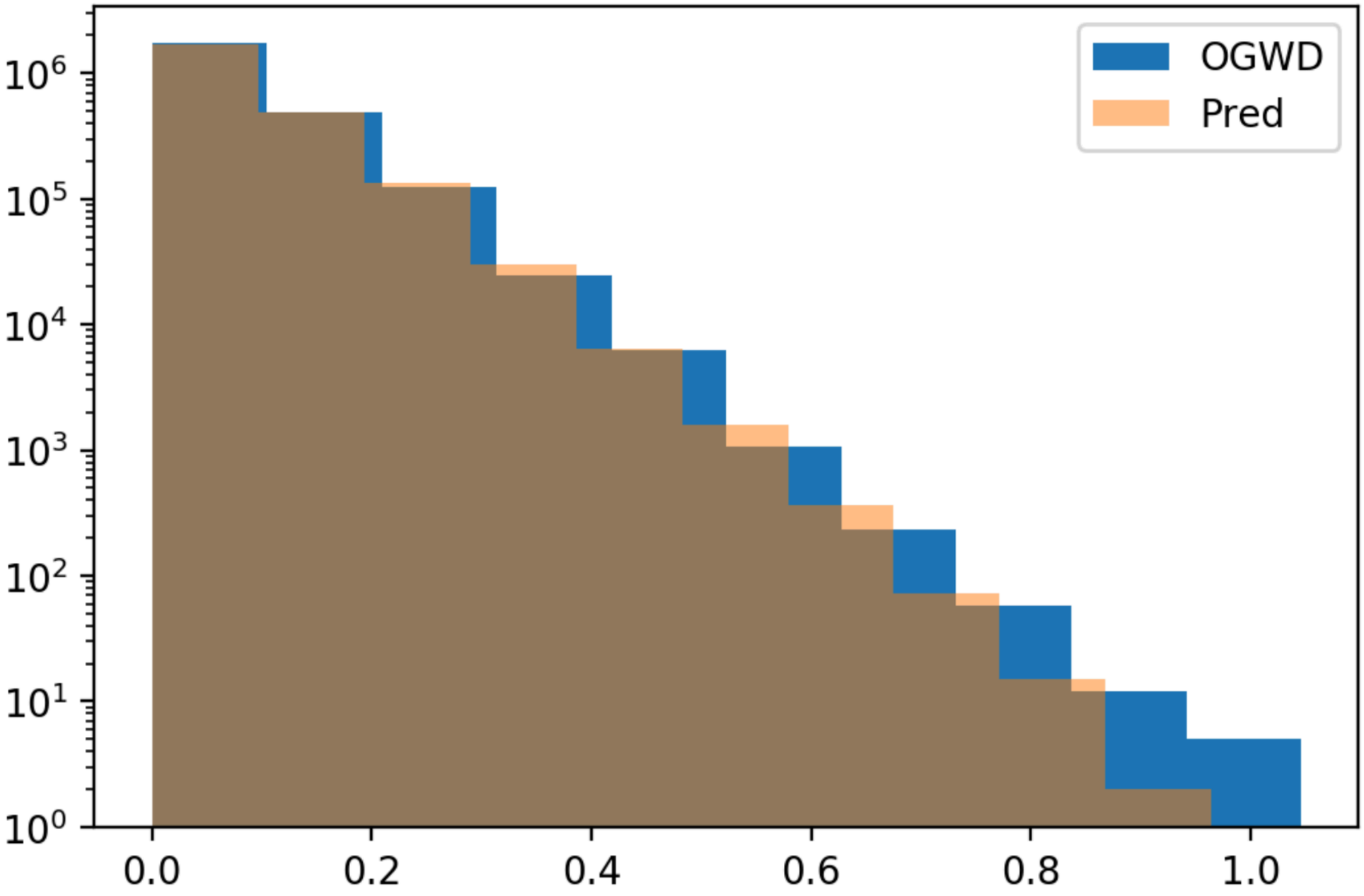
1e+05 mseswish RMSE : 0.336



Are the rare events more rare?



How about predicting extremes?



OGWD scheme predicts extremes but not in the right places.

Networks unable to learn importance of surface parameters.