Application of dimensional reduction in the training of Machine Learning-based emulators for biogeochemical downscaling of the Northeast Pacific Ocean

Albert J. Hermann

University of Washington

Cooperative Institute for Climate Ocean and Ecosystem Studies (CICOES)

In collaboration with other non-federal colleagues:

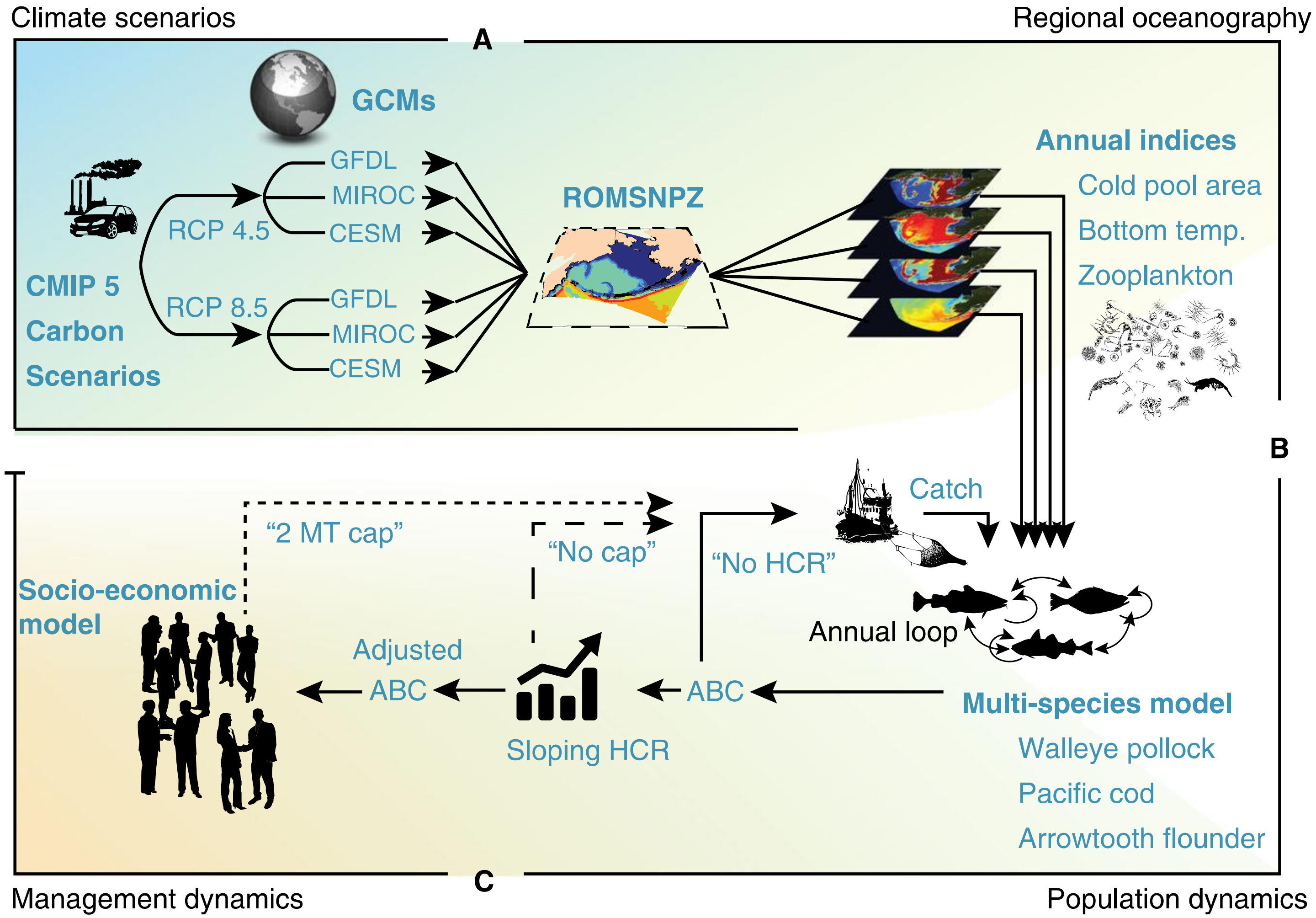
Vivek Seelanki (CICOES), Wei Cheng (CICOES), Kate Hedstrom (UAF)

PICES 2025 Annual Meeting Session S10

Downscaled climate projections used in fisheries management strategy evaluation

Ideally want really big ensembles for management applications

These can be very costly!



(Holsman et al 2020)

CEFI NEP10k Domain and Hindcast Configuration

Bathymetry (right, in meters): GEBCO 2020

Temporal Extent: 1993-2019 (27 years)

Atmospheric Forcing: ***JRA-55***

Tidal Forcing: TPXO

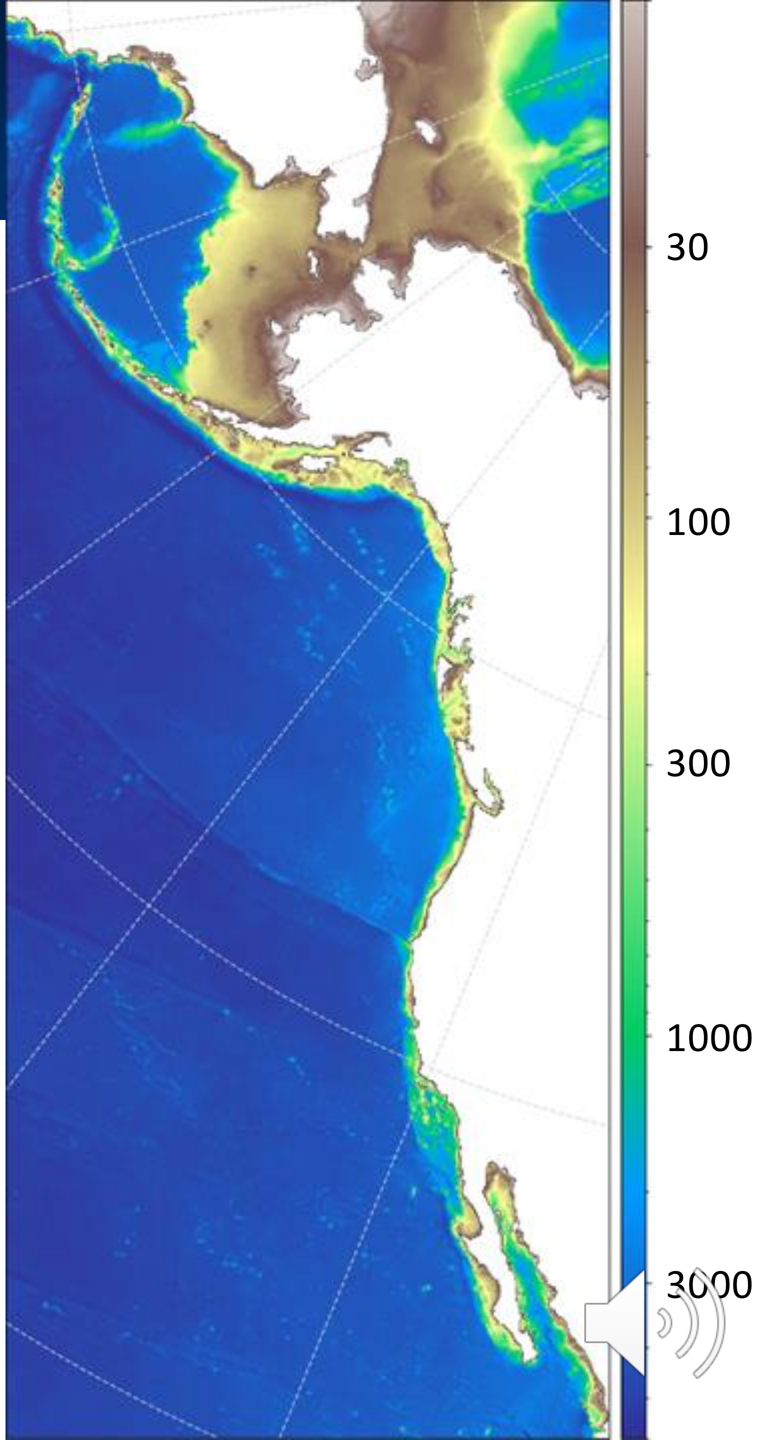
<u>River Forcing</u>

Freshwater: GloFAS, Beamer et al., (2016; GoA)

<u>Initial and Boundary Conditions</u>

Ocean Physics: GLORYS12

(modified slide from L. Drenkard)



Surrogate Modeling

- MOTIVATION: Regional models are computationally expensive!
- Output from a complex model can be used to train a ***surrogate*** which compactly approximates the behavior of the full system
- The use of a compact surrogate allows a broader range of model experiments, e.g.:
  - Quantify sensitivities to forcing and parameters
  - Broaden ensemble of predictions
- Here, we explore the use of ***Machine Learning*** to construct a 3D surrogate ("emulator") for a regional NEP model based on MOM6
- Machine Learning can include EOF analysis (a form of "unsupervised learning")

EOFs have a long history of use to identify dominant geospatial patterns and their time variation

- Examples include:
  - ENSO
  - The Pacific Decadal Oscillation

- There are many others
  - Some (e.g. NPGO) are *not* the leading mode of variability!

EOF analysis is based on *Singular Value Decomposition* of a matrix

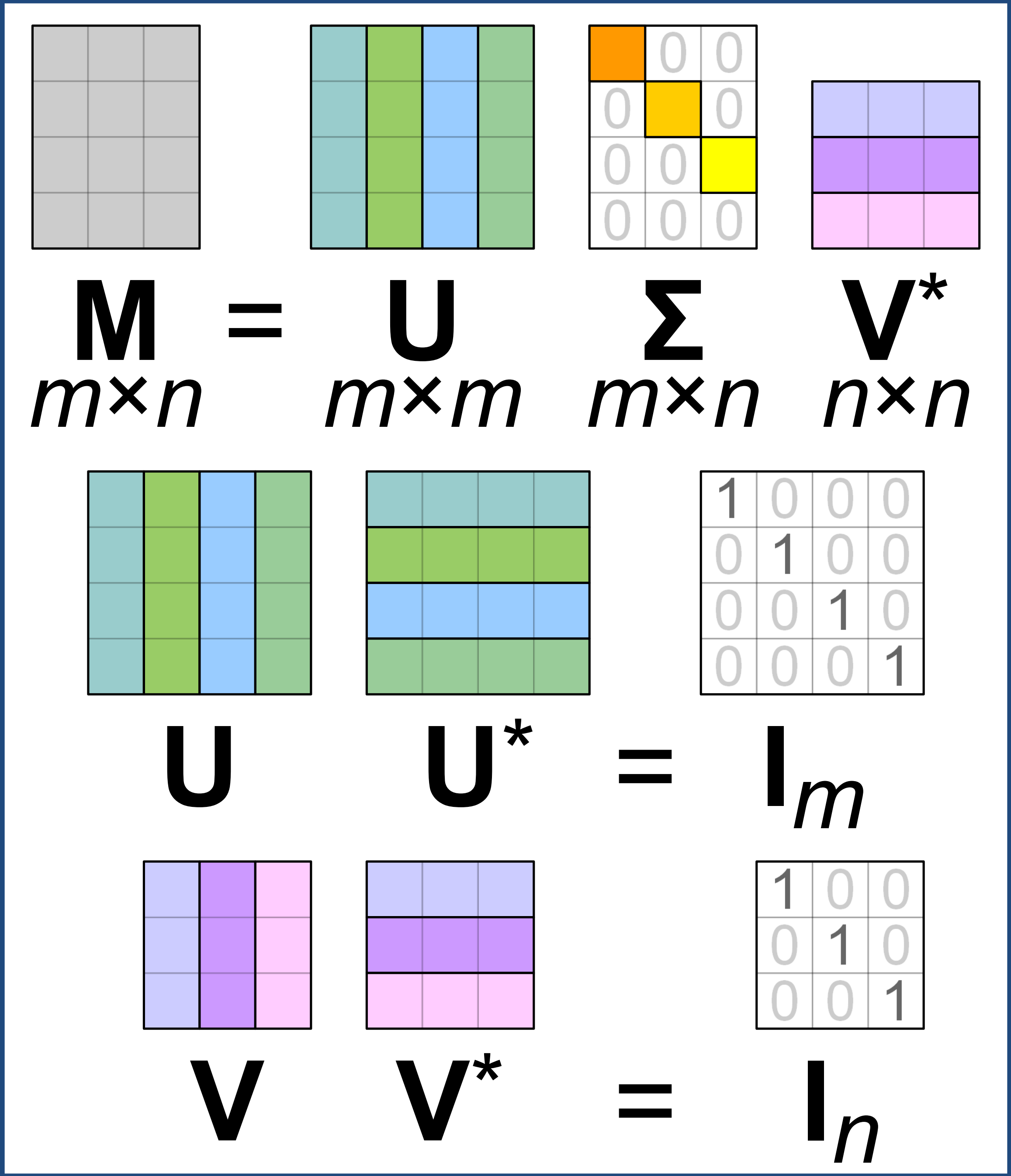
The two dimensions can really be anything:

space x time (univariate EOFs)

variable x time (Principal Components)

variable/space x time (multivariate EOFs)

FIGURE: By Cmglee - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=67853297



EOFs *can* represent signals propagating through space or across different variables

EOF decomposition (which is just Singular Value Decomposition of a matrix) typically uses a collection of time series at multiple locations:

$$V(x,t) = X1(x)*T1(t) + X2(x)*T2(t) + ...$$

- The SVD-based calculation of these modes just sees a collection of time series, which can include multiple variables as well as multiple locations.
- EOFs can represent a propagating signal (across space/time/variables) according to the algebraic equivalence:

$$\sin(kx - \omega t) = \sin(kx)*\cos(\omega t) - \cos(kx)*\sin(\omega t)$$

- Any rearrangement of the time series does not affect the resulting X and T! hence EOFs can even represent propagating signals with spatially variable phase speeds.

Advantages and disadvantages of EOFs

- Advantages
  - Allow complete reconstruction of original data
  - Orthogonal spatial *and* temporal modes
- Disadvantages
  - Orthogonal spatial *and* temporal modes requirement may obscure simple signals (note Fourier decomposition does not require this)
  - Chaotic, small-scale features will not be well captured (because spatially/temporarily irregular)
  - Need a significant number of independent realizations of a pattern to get significant EOFs
- Other methods exist! Many are now used in Machine Learning

Principal Component analysis by itself can be used for hybrid dynamical-statistical downscaling

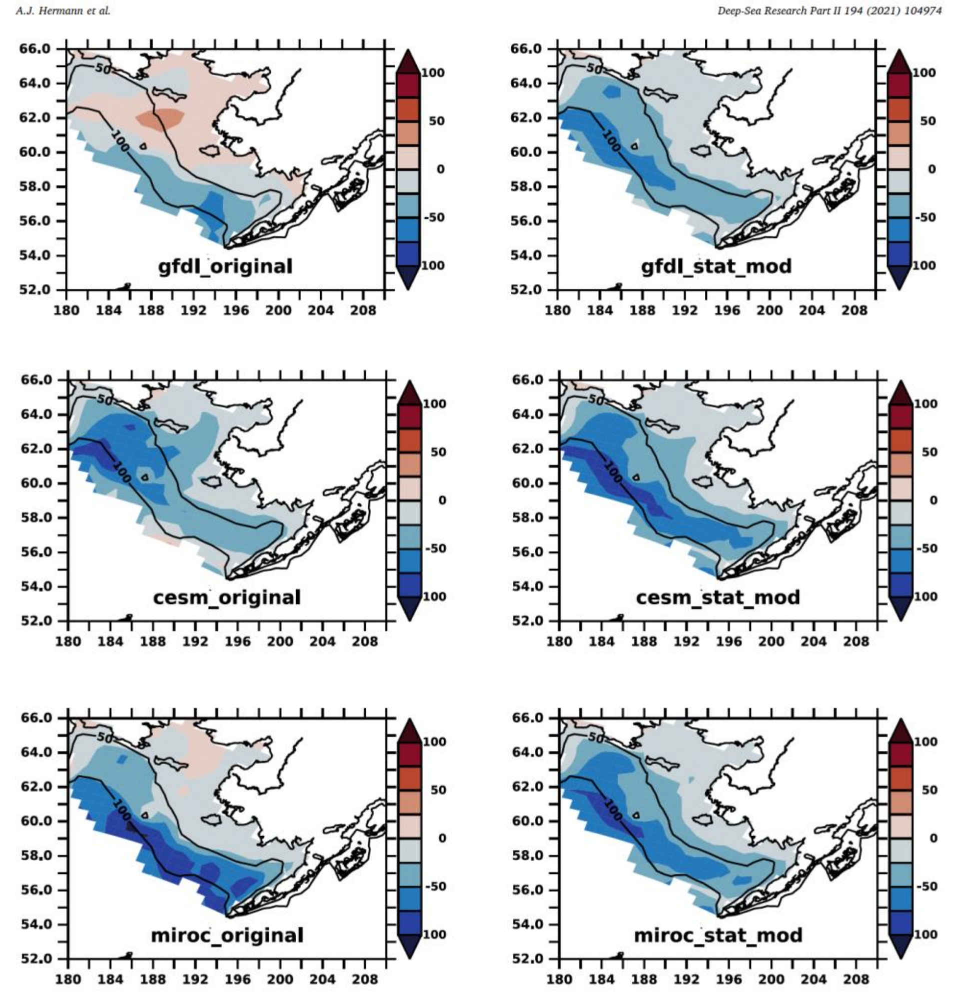
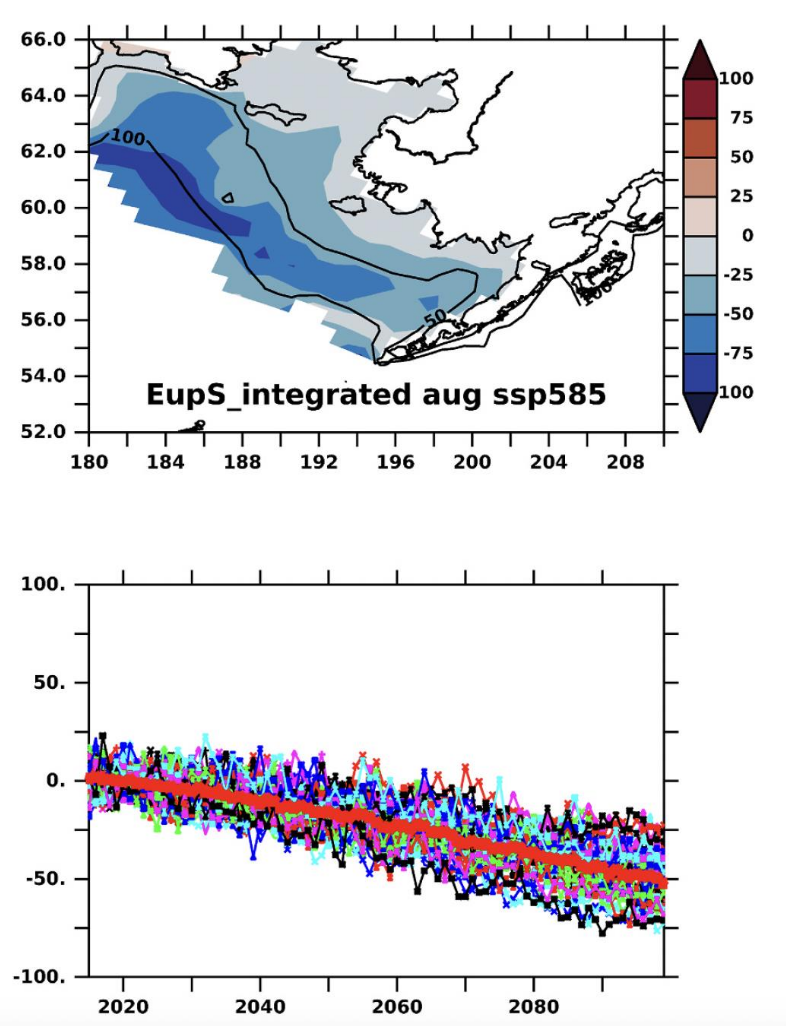
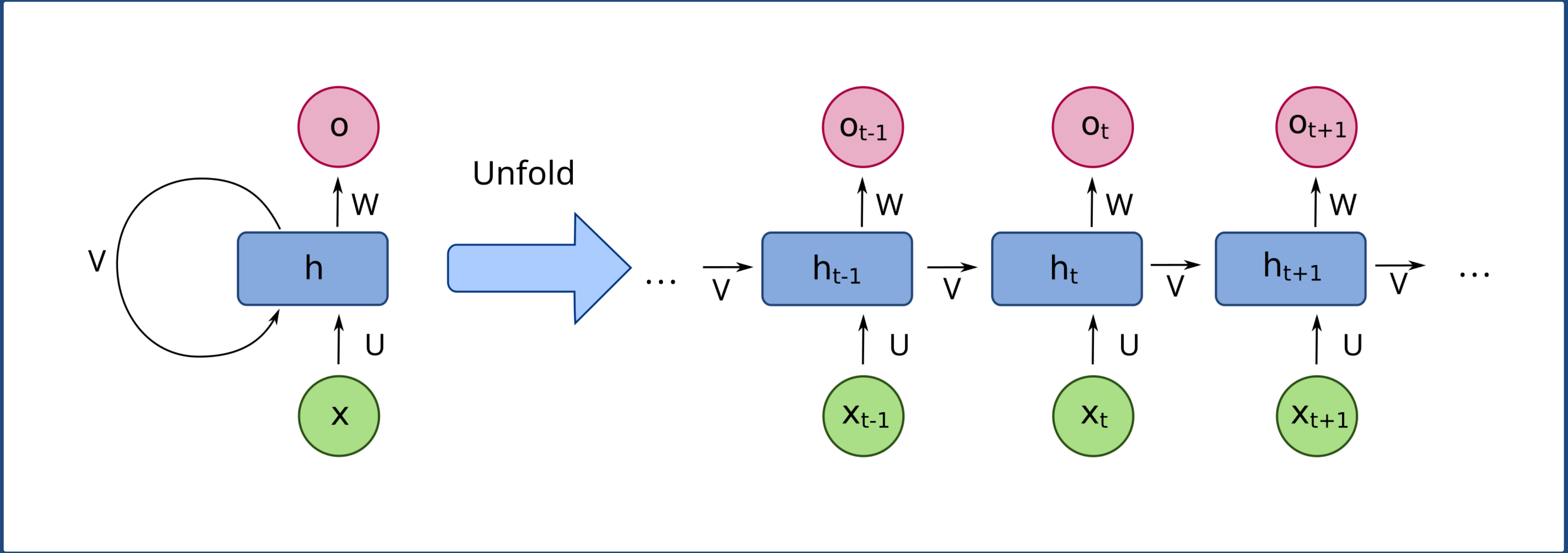


Fig. 21. As in Fig. 19, for July depth-integrated shelf Euphausiids (EupS, mg C m^{-2}).

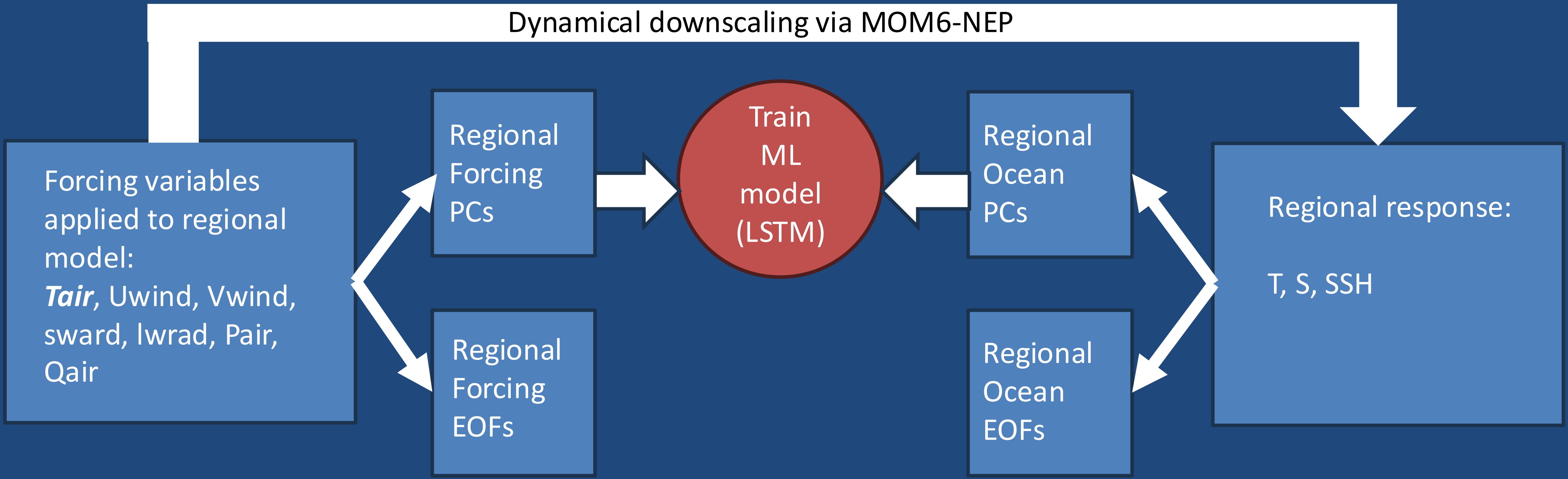


Hermann et al. 2021, Deep-Sea Res II

Recurrent Neural Networks can be used to relate two sets of time series (LSTM is a variant of this)

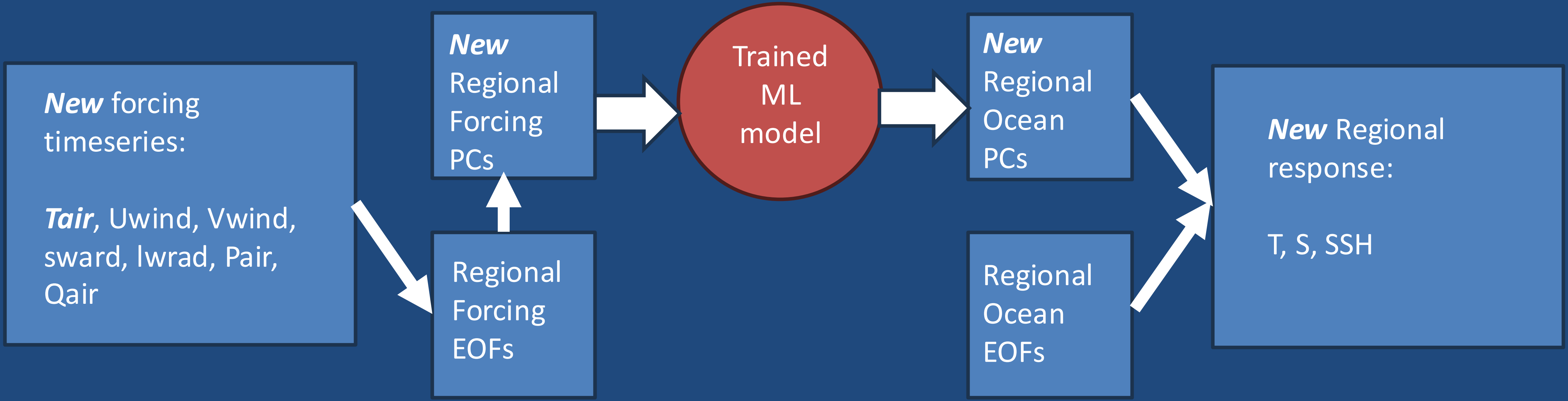


We dynamically downscale, calculate forcing and response EOFs of monthly anomalies, then train the ML model to relate the PCs



- Include the past 12 months of forcing for training and emulation
- Include top 20 PCs of each forcing (2D) and top 20 PCs of each response variable (3D)
- Use 400 LSTM “neurons” in the LSTM
- Optimization target for each “training session” can be a single PC of a single regional response variab *can train all variables/modes simultaneously*

We then project new forcing sets onto the regional forcing EOFs and use the ML model to emulate the regional response to that ***new forcing***



In Machine Learning terms: we are using Principal Component Analysis as the *Encoder/Decoder* bracketing the LSTM

Method details and timing

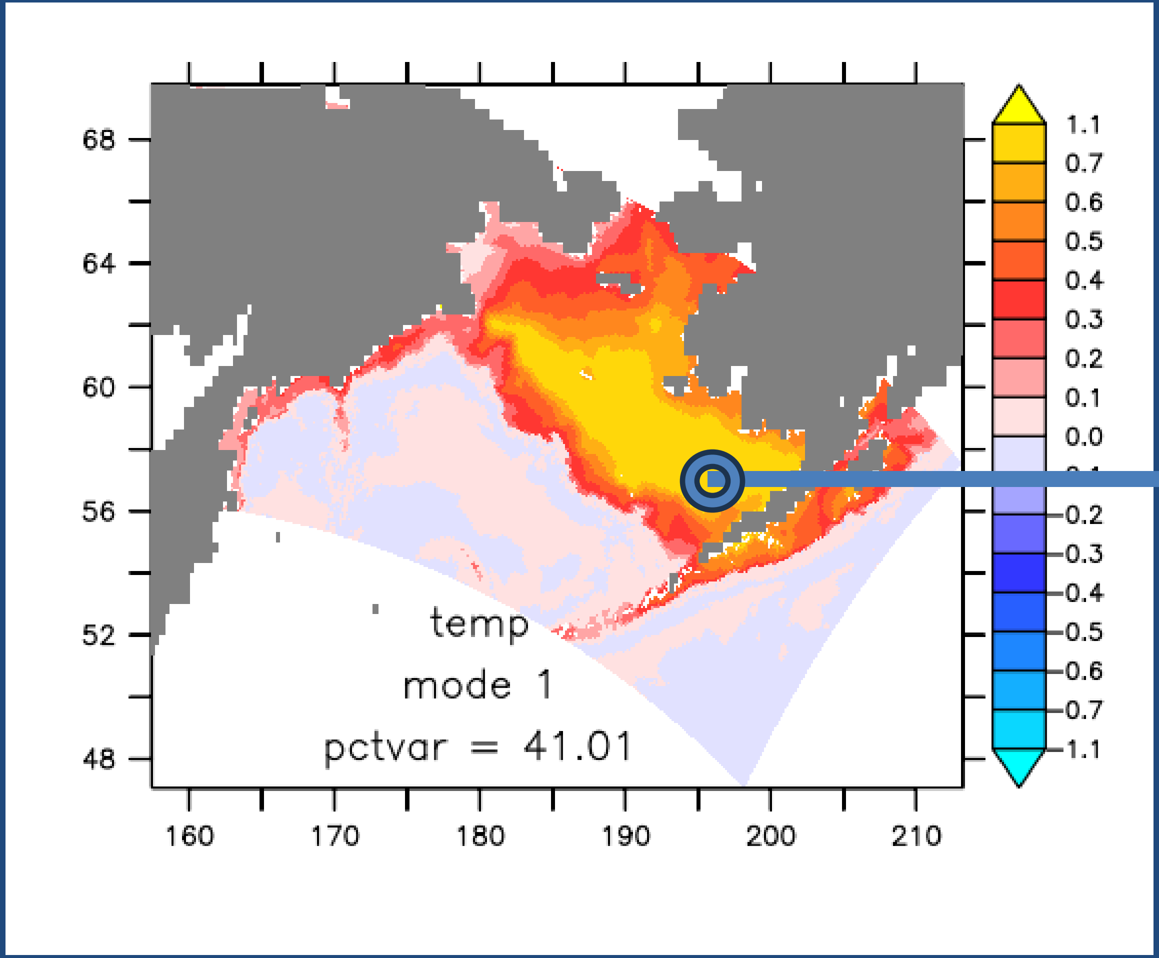
- *Train* using 1993-2009 series; *Test* using 2010-2018
- Timing statistics
  – Run dynamical model 1993-2018 (~200 cpu-days)
  – ***train*** with a hindcast of 1993-2009 (~240 cpu-sec)
  – ***test*** with a hindcast of 2010-2018 (~***1 cpu-sec***)

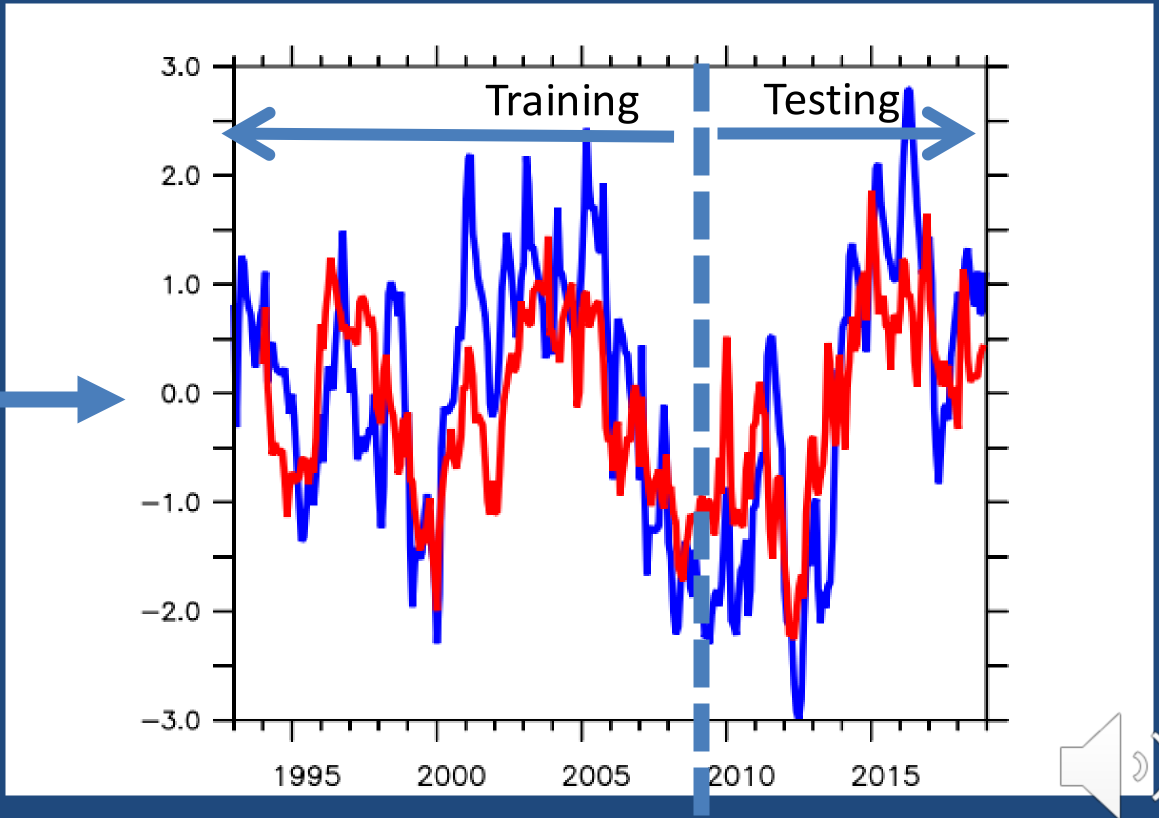
Temperature (deg C) results for the BERING SEA SHELF

Left panel: leading mode ***3D*** EOF of temperature (values at the sea floor)

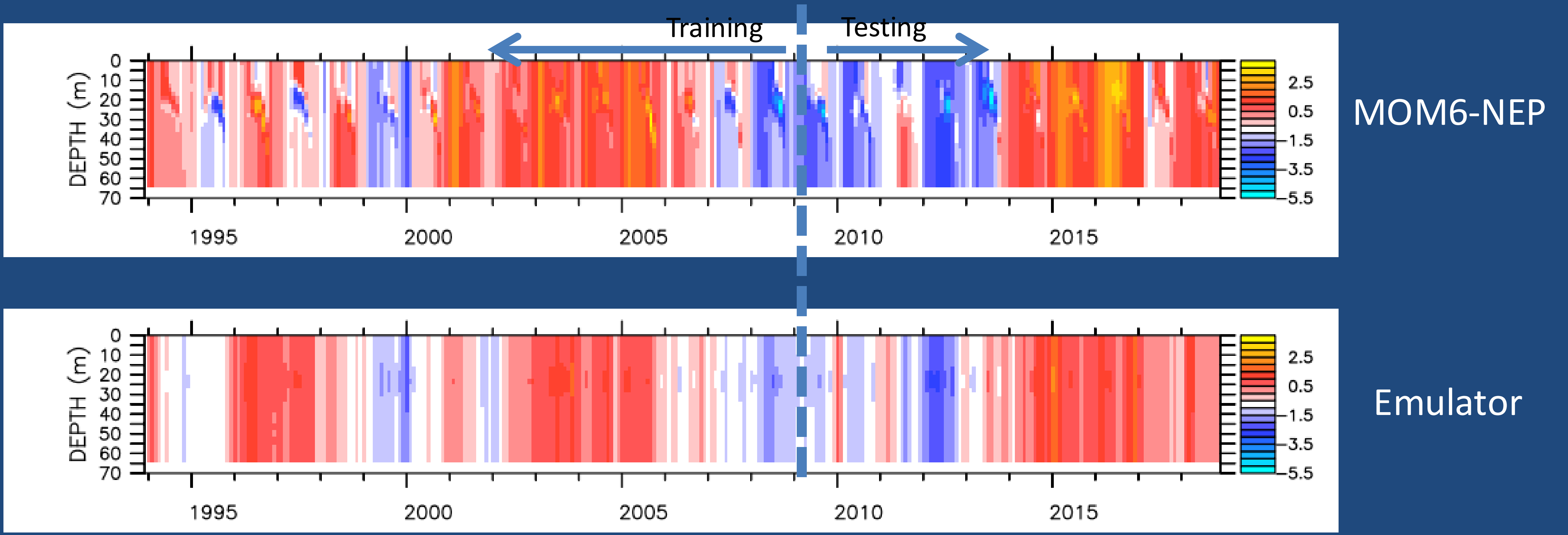
Right panel: monthly anomalies of bottom temperature at mid-shelf mooring "M2"

(Blue = MOM6-NEP; Red = Emulator, ***summed over all EOF modes***)



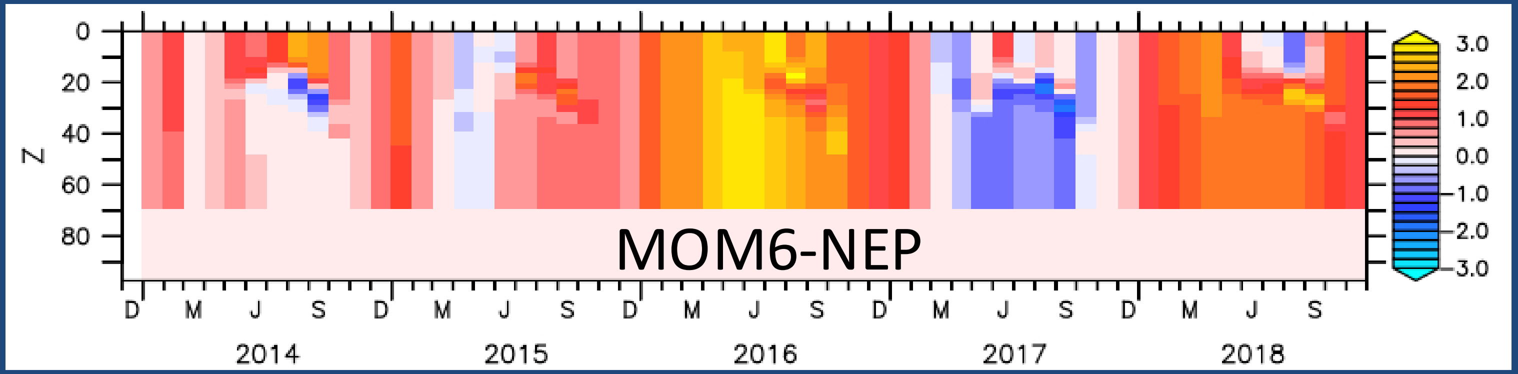


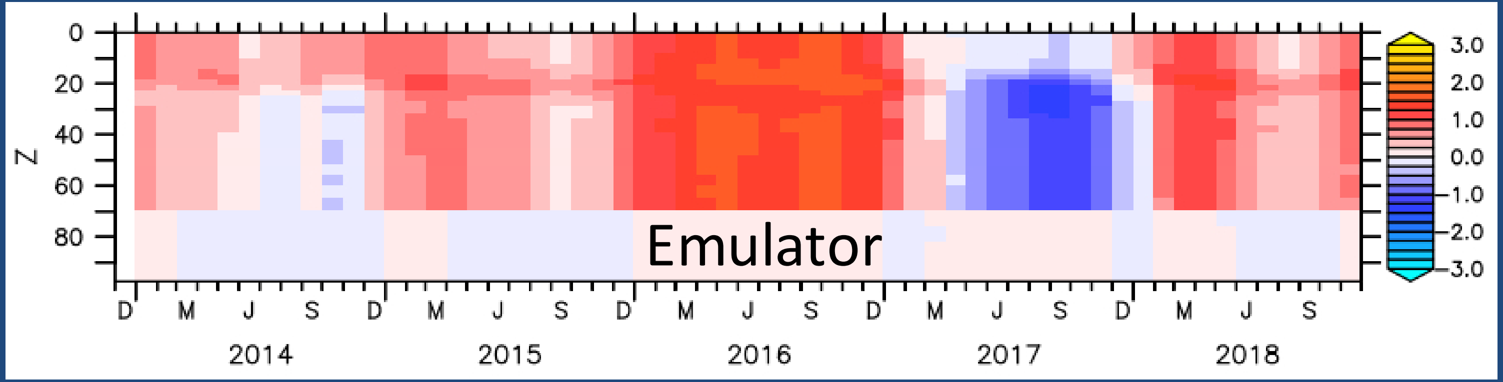
Monthly anomaly temperature profile at M2



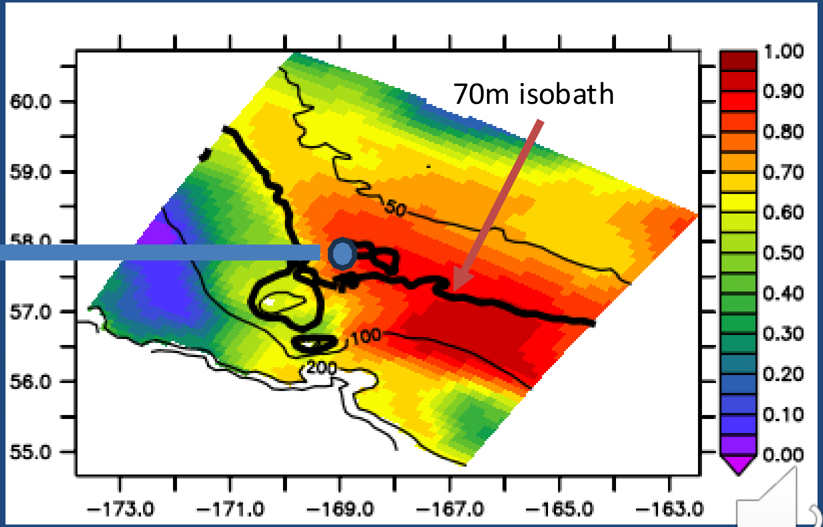
New results using "direct" method w/o EOFs better at vertical gradients but skill is more "local"

Monthly T *anomaly* profiles at M4 (validation period only)

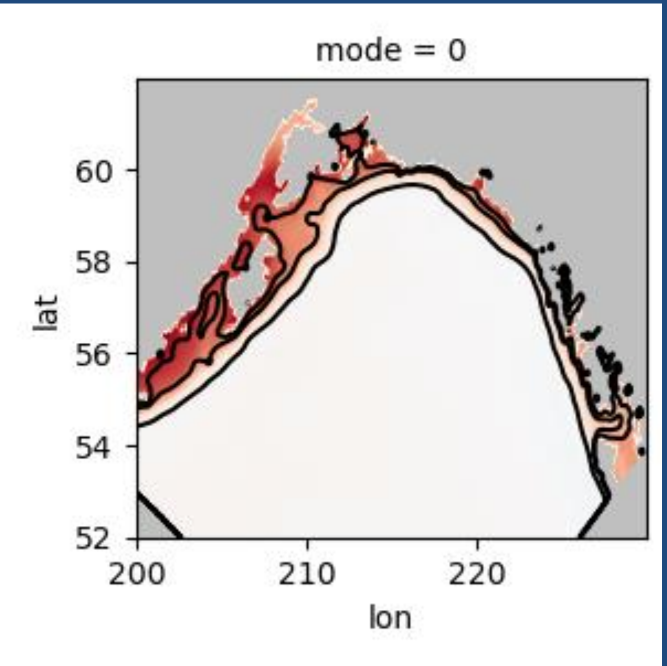


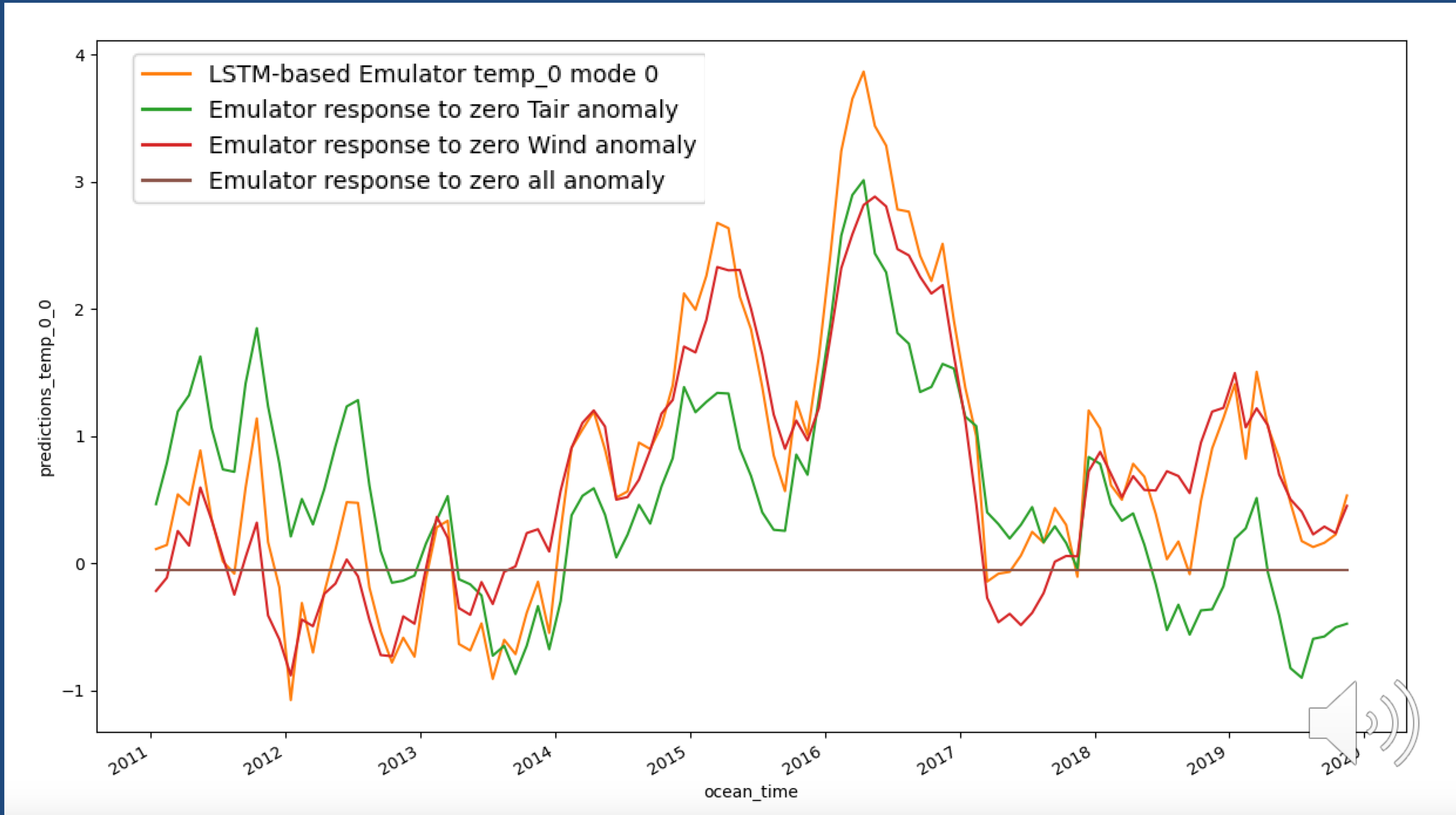


Bottom T Correlation MOM6-NEP vs. Emulator (red => r = 1.0)



Emulators can be used for *sensitivity analysis* (ROMS example): base emulator (orange), no Tair (green), no winds (red)



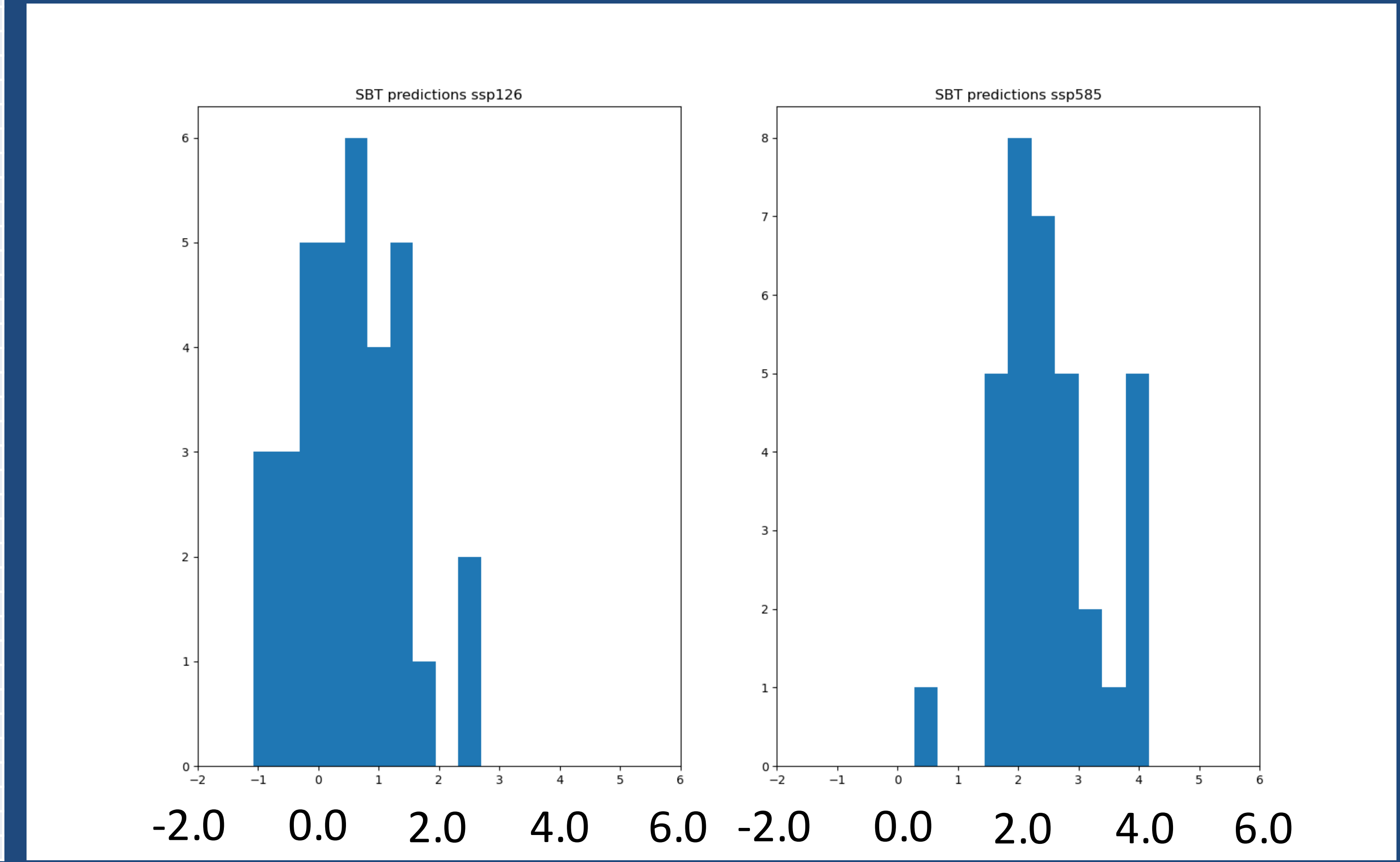


Feed a big CMIP6 ensemble of monthly air temperatures into the trained model and compare SBT under ssp126 vs ssp585

- ACCESS-CM2_ssp126_r1i1p1f1_gn
- ACCESS-CM2_ssp126_r2i1p1f1_gn
- ACCESS-ESM1-5_ssp126_r1i1p1f1_gn
- ACCESS-ESM1-5_ssp126_r2i1p1f1_gn
- AWI-CM-1-1-MR_ssp126_r1i1p1f1_gn
- BCC-CSM2-MR_ssp126_r1i1p1f1_gn
- CAMS-CSM1-0_ssp126_r1i1p1f1_gn
- CAMS-CSM1-0_ssp126_r2i1p1f1_gn
- CanESM5_ssp126_r1i1p1f1_gn
- CanESM5_ssp126_r2i1p1f1_gn
- CAS-ESM2-0_ssp126_r1i1p1f1_gn
- CESM2-WACCM_ssp126_r1i1p1f1_gn
- CIESM_ssp126_r1i1p1f1_gr
- CMCC-CM2-SR5_ssp126_r1i1p1f1_gn
- CMCC-ESM2_ssp126_r1i1p1f1_gn
- CNRM-CM6-1-HR_ssp126_r1i1p1f2_gr
- CNRM-CM6-1_ssp126_r1i1p1f2_gr
- CNRM-ESM2-1_ssp126_r1i1p1f2_gr
- EC-Earth3_ssp126_r1i1p1f1_gr
- EC-Earth3-Veg-LR_ssp126_r1i1p1f1_gr
- EC-Earth3-Veg-LR_ssp126_r2i1p1f1_gr
- EC-Earth3-Veg_ssp126_r1i1p1f1_gr
- EC-Earth3-Veg_ssp126_r2i1p1f1_gr
- FGOALS-f3-L_ssp126_r1i1p1f1_gr
- FGOALS-g3_ssp126_r1i1p1f1_gn
- FGOALS-g3_ssp126_r2i1p1f1_gn
- GFDL-ESM4_ssp126_r1i1p1f1_gr1
- GISS-E2-1-G_ssp126_r1i1p1f2_gn
- IITM-ESM_ssp126_r1i1p1f1_gn
- KIOST-ESM_ssp126_r1i1p1f1_gr1
- MCM-UA-1-0_ssp126_r1i1p1f2_gn
- MIROC-ES2L_ssp126_r1i1p1f2_gn
- MPI-ESM1-2-HR_ssp126_r1i1p1f1_gn
- MPI-ESM1-2-HR_ssp126_r2i1p1f1_gn
- MRI-ESM2-0_ssp126_r1i1p1f1_gn
- NESM3_ssp126_r1i1p1f1_gn
- NESM3_ssp126_r2i1p1f1_gn
- NorESM2-LM_ssp126_r1i1p1f1_gn
- NorESM2-MM_ssp126_r1i1p1f1_gn
- TaiESM1_ssp126_r1i1p1f1_gn
- UKESM1-0-LL_ssp126_r1i1p1f2_gn

Histograms of change near Shelikof Strait July 2015->2100

ssp126 ssp585



- ACCESS-CM2_ssp585_r1i1p1f1_gn
- ACCESS-CM2_ssp585_r2i1p1f1_gn
- ACCESS-ESM1-5_ssp585_r1i1p1f1_gn
- ACCESS-ESM1-5_ssp585_r2i1p1f1_gn
- AWI-CM-1-1-MR_ssp585_r1i1p1f1_gn
- BCC-CSM2-MR_ssp585_r1i1p1f1_gn
- CAMS-CSM1-0_ssp585_r1i1p1f1_gn
- CAMS-CSM1-0_ssp585_r2i1p1f1_gn
- CanESM5_ssp585_r1i1p1f1_gn
- CanESM5_ssp585_r2i1p1f1_gn
- CAS-ESM2-0_ssp585_r1i1p1f1_gn
- CESM2-WACCM_ssp585_r1i1p1f1_gn
- CESM2-WACCM_ssp585_r2i1p1f1_gn
- CIESM_ssp585_r1i1p1f1_gr
- CMCC-CM2-SR5_ssp585_r1i1p1f1_gn
- CMCC-ESM2_ssp585_r1i1p1f1_gn
- CNRM-CM6-1-HR_ssp585_r1i1p1f2_gr
- CNRM-CM6-1_ssp585_r1i1p1f2_gr
- CNRM-ESM2-1_ssp585_r1i1p1f2_gr
- EC-Earth3_ssp585_r1i1p1f1_gr
- EC-Earth3-Veg-LR_ssp585_r1i1p1f1_gr
- EC-Earth3-Veg-LR_ssp585_r2i1p1f1_gr
- EC-Earth3-Veg_ssp585_r1i1p1f1_gr
- EC-Earth3-Veg_ssp585_r2i1p1f1_gr
- FGOALS-f3-L_ssp585_r1i1p1f1_gr
- FGOALS-g3_ssp585_r1i1p1f1_gn
- FGOALS-g3_ssp585_r2i1p1f1_gn
- GFDL-ESM4_ssp585_r1i1p1f1_gr1
- GISS-E2-1-G_ssp585_r1i1p1f2_gn
- IITM-ESM_ssp585_r1i1p1f1_gn
- KIOST-ESM_ssp585_r1i1p1f1_gr1
- MCM-UA-1-0_ssp585_r1i1p1f2_gn
- MIROC-ES2L_ssp585_r1i1p1f2_gn
- MPI-ESM1-2-HR_ssp585_r1i1p1f1_gn
- MPI-ESM1-2-HR_ssp585_r2i1p1f1_gn
- MRI-ESM2-0_ssp585_r1i1p1f1_gn
- NESM3_ssp585_r1i1p1f1_gn
- NESM3_ssp585_r2i1p1f1_gn
- NorESM2-LM_ssp585_r1i1p1f1
- NorESM2-MM_ssp585_r1i1p1f1
- TaiESM1_ssp585_r1i1p1f1_gn
- UKESM1-0-LL_ssp585_r1i1p1f2_gn

Conclusions and next steps

- Machine Learning methods show promise as ***fast*** downscaling model emulators
- After training, the broad-scale regional ocean response can be largely emulated using only atmospheric forcing
- Some spatial details of the regional ocean were lost using EOFs, but some broad spatial patterns were hard to capture without them!
- Next steps:
  - Explore training of the ML model using raw atmospheric fields (w/o EOF reduction) but retain EOFs for dimensional reduction of the oceanic response and utilize more modes (to get more of the total variance).