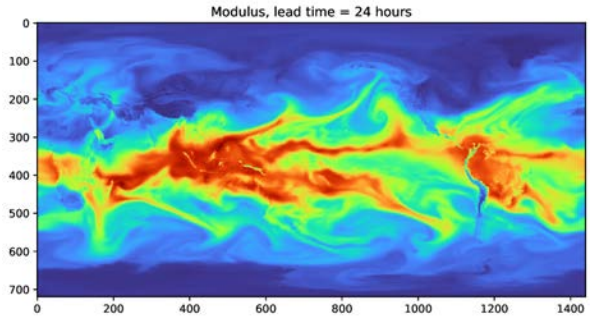
Hewlett Packard
Enterprise

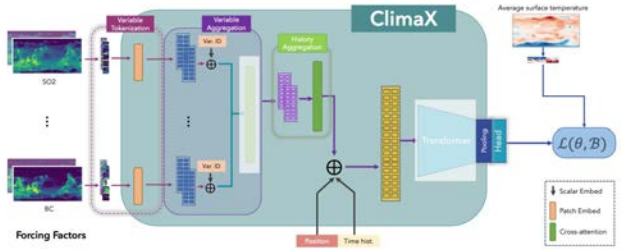
An online, continuous learning framework for training surrogates of ocean models

Andrew Shao [HPE Canada]
Alistair Adcroft [GFDL/Princeton]

September 10, 2024
Ocean Model Development, Data-driven Parameterizations, and
Machine Learning in Ocean Models of the Earth System Workshop

Surrogate modelling in climate/weather





1. Subgrid-scale surrogates

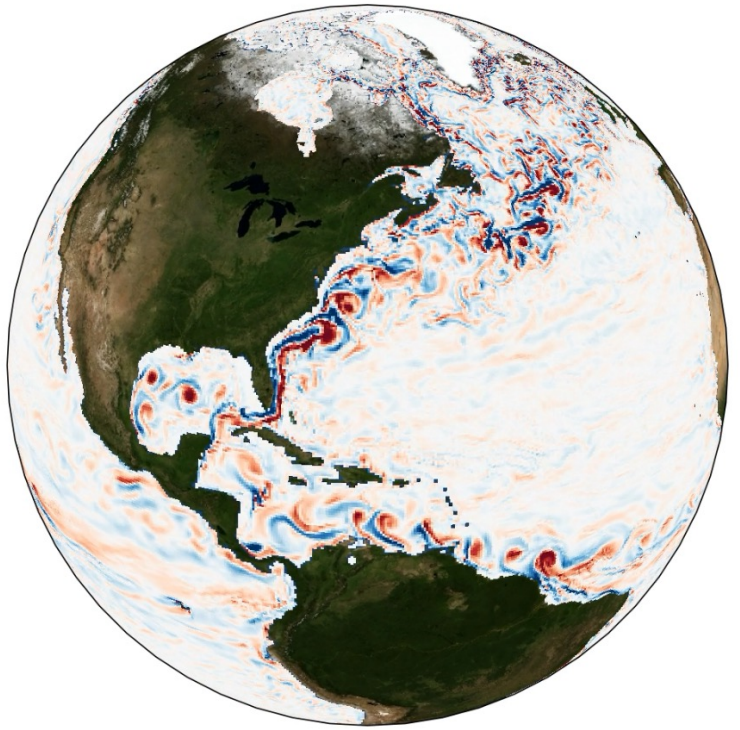
- Neural network as parameterizations
- Examples
  - Eddy kinetic energy, backscatter

2. Full Surrogate:

- Train neural network to fully replace a simulation
- Examples
  - FourCastNet, GraphCast, ClimaX, FuXi, Pangu-Waeather

3. Numerical surrogate

- Replace a numerical scheme with AI model
- Examples
  - 1D advection operator (atmosphere)
  - Barotropic solver (Iuri Gorenstein, Yesterday)



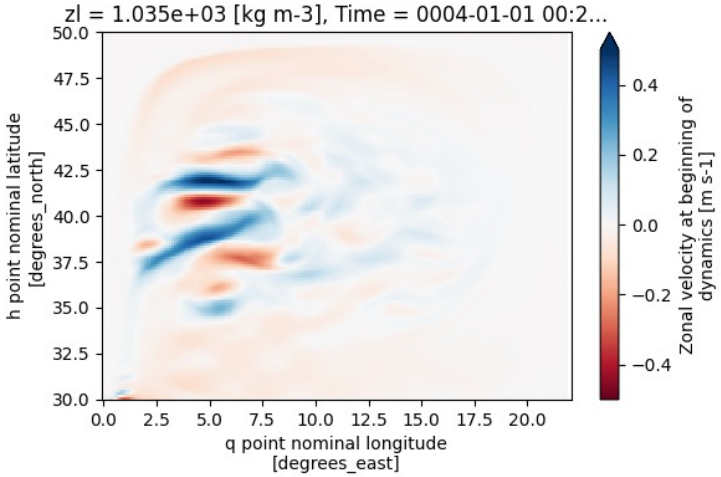
Guillaumin and Zanna [2021]
CNN Backscatter in OM4 ¼-degree (1 year)

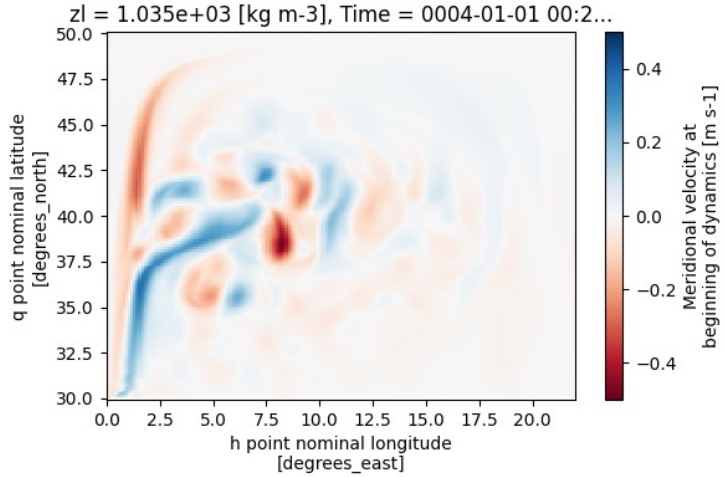
Why numerical surrogate modelling?

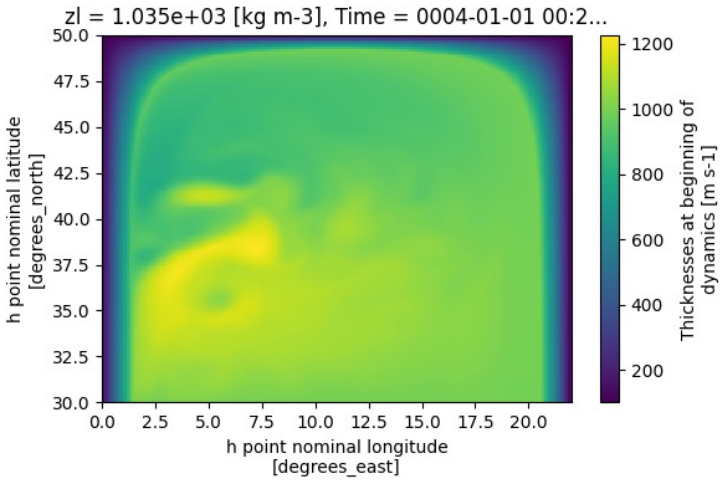
1. **Acceleration [Focus for this presentation]**
   - Neural networks run on GPUs, potential for orders of magnitude levels of improvement
2. **Accuracy**
   - Estimate spatial/temporal truncation errors [Fabricio Rodrigues Lapolli, yesterday)
   - Act as a pre-conditioner for iterative solvers
3. **Adjoint surrogate**
   - Neural networks automatically have the gradient of the operator (used in training)
4. **Keeps all other "machinery" in place**
   - Can continue to rely on most diagnostics
   - Coupling to physical and BGC models
   - Virtuous cycle between numerics development and surrogate
5. **Model can be kept small**
   - Hard work was done by the deriver/implementer of numerical method

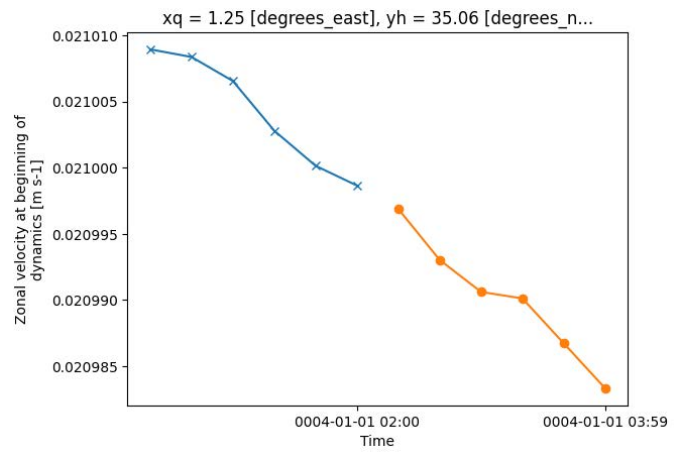
Model/AI setup

- **MOM6 double-gyre:**
  - Baroclinic timestep: 900s
  - Resolution: 1/8-degree
  - Vertical: 2 layers
- **Neural network:**
  - Resnet18 + GRU layer
  - Input: 6 snapshots of u, v, h (2 layers)
  - Output: 6 snapshots of uh, vh
- **Main Idea:**
  - **Predict the mass transports, not model state**
    - Flux-form guarantees conservation!
    - Fluxes can also be used for tracers transport
- **“Easy” problem**
  - Original calculation just uses a 5-point stencil for every u-velocity
- **“Hard” Problem**
  - Split barotropic/baroclinic
  - Predicting multiple time levels requires a complex, wider stencil



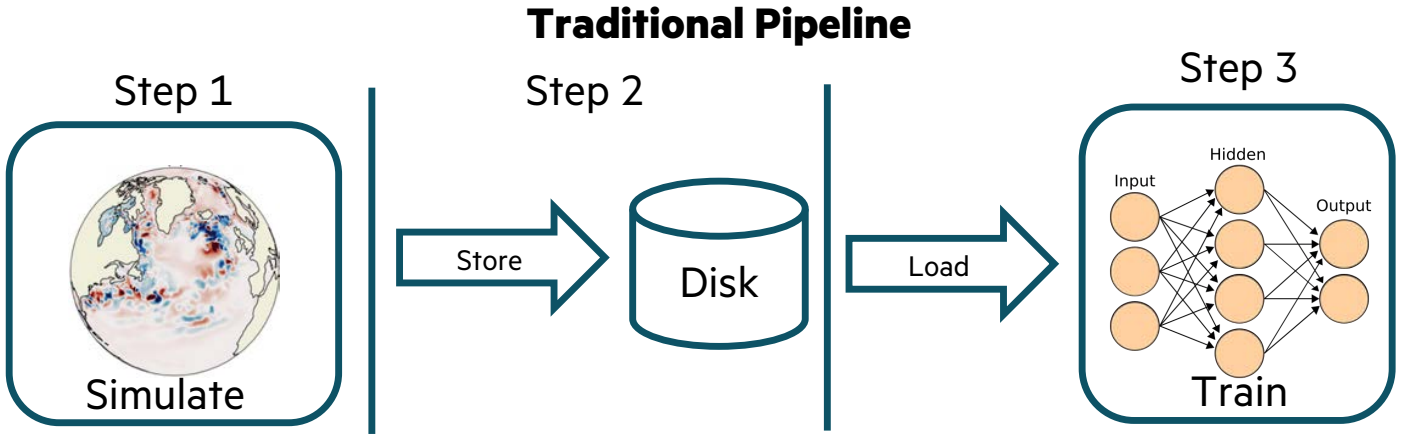


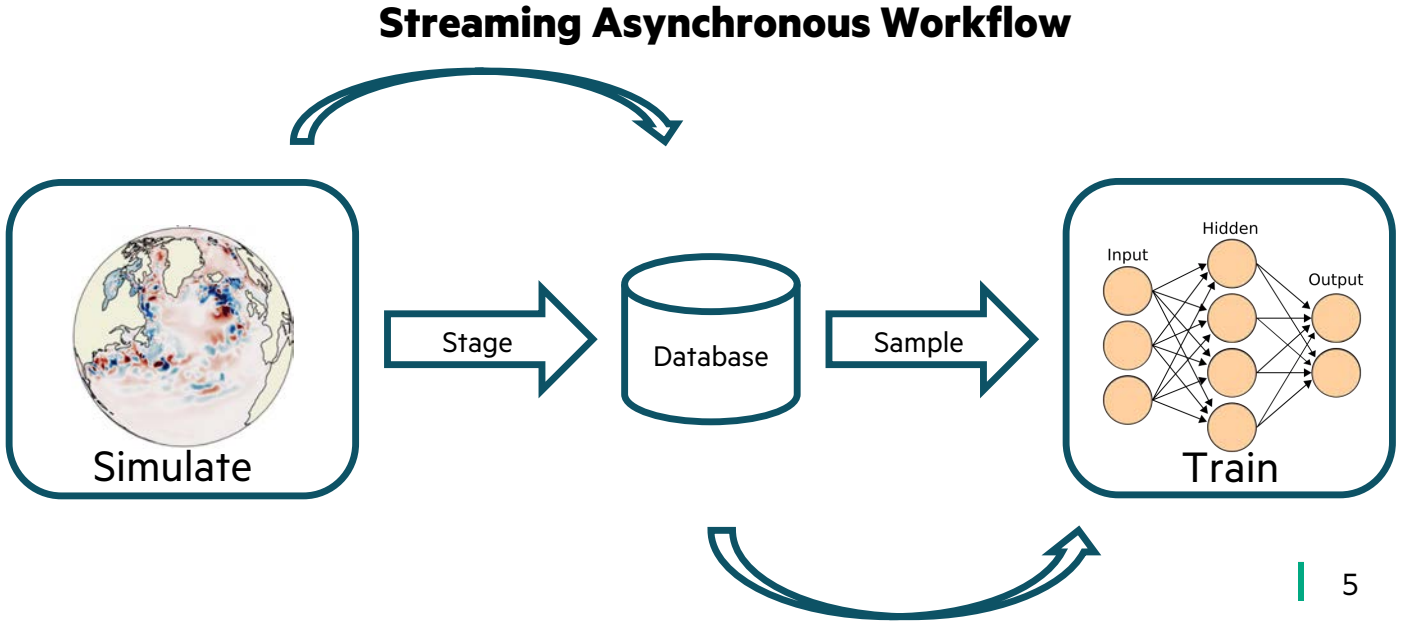




Data volume is a fundamental problem for numerical surrogates

- **Numerical surrogates need timestep-level data**
  - (1PB/model year ¼-degree model)
- **Generally want to map a function-space to another function space**
  - Oversampling a portion of the sample space biases model
- **Questions:**
  - How do you store this amount of data?
  - How do you train on this amount of data?
- **Solution:**
  - Sample the data in an 'intelligent' way
  - Train surrogate in a streaming manner





HPE's SmartSim to enable the workflow

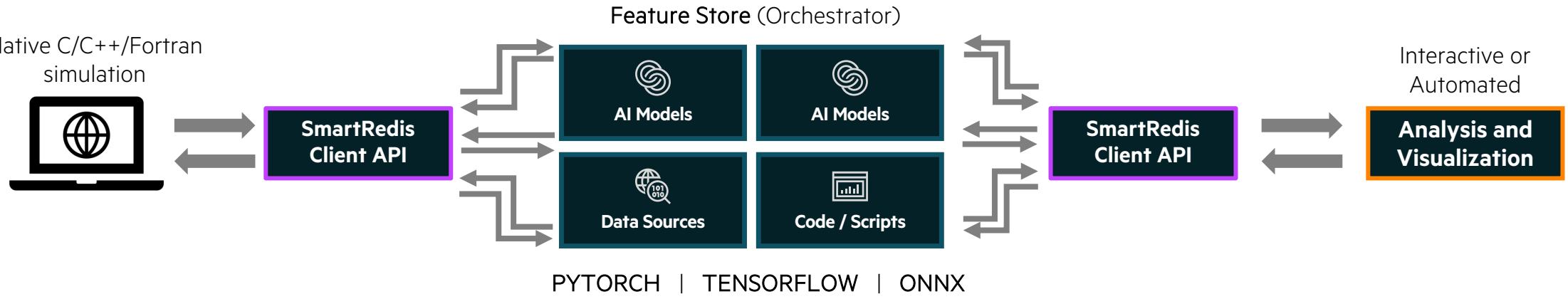
SmartSim is an open-source library

- Providing a loose-coupling philosophy for combining HPC & AI

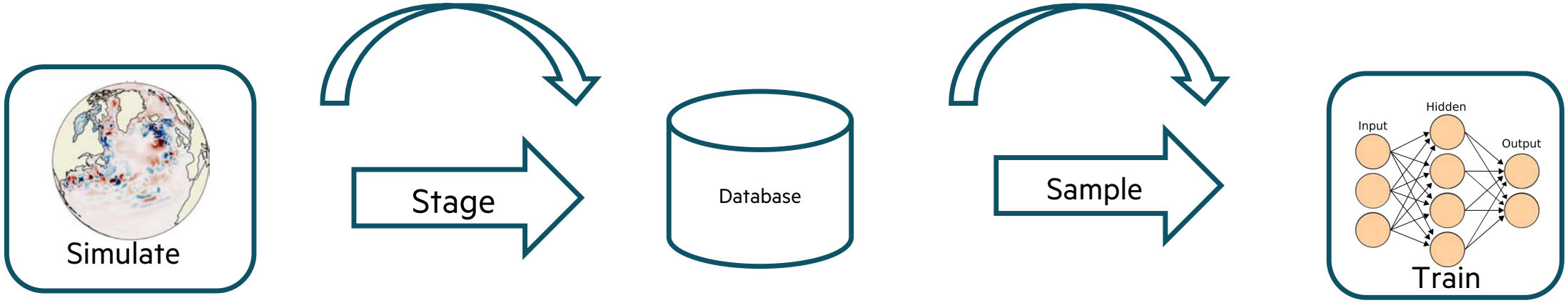
SmartSim allows scientists to create complex workflows, with simulations and machine learning components producing and consuming data

- Call Machine Learning (ML) inference in existing Fortran/C/C++ simulations
- Exchange data between C, C++, Fortran, and Python applications
- Train ML models online and make predictions using TensorFlow, PyTorch, and ONNX
- Analyze data streamed from HPC applications while they are running

Recent Paper: Combining machine learning with computational fluid dynamics using OpenFOAM and SmartSim



Setting up the asynchronous, workflow



MOM6

- Store 12 dynamics timesteps
  - u, v, h at beginning of timestep
  - uh, vh at end of timestep
- Stage in database
  - Sent using SmartRedis
  - Stored in SmartSim database

Intelligent Sampler

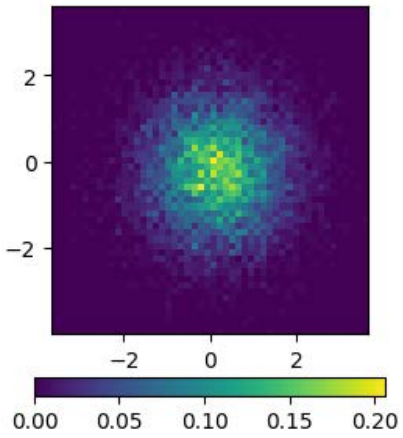
- Polls database for 8 new datasets
- Performs statistical comparison to accept/reject new samples
- Stores downsampled data in database
- Delete original data

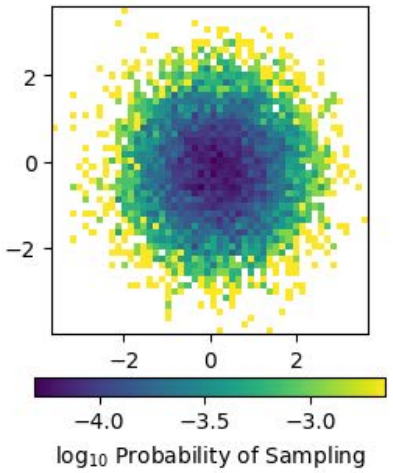
Trainer

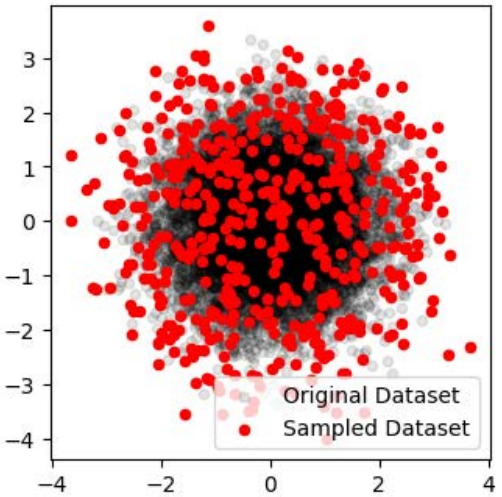
- Infinite training loop
- Every *N* training epochs, update dataloader to check for new data

Intelligent sampling for point-by-point predictions

- Problem: AI susceptible to sampling bias
  - With most PDEs, no part of the solution space is more “valid”
  - Naïve training of AI models leads to
    - Fixating on the well-sampled parts of the domain
    - Ignoring the outliers
  - Example: Ocean generally is energetic near the surface, more quiescent in the interior
    - “Average” behaviour (by volume) of the ocean is relatively sedentary
    - Uniformly sampling of the ocean creates a dataset biased towards the low end
- Solution: Inversely sample the data to promote uniform sampling

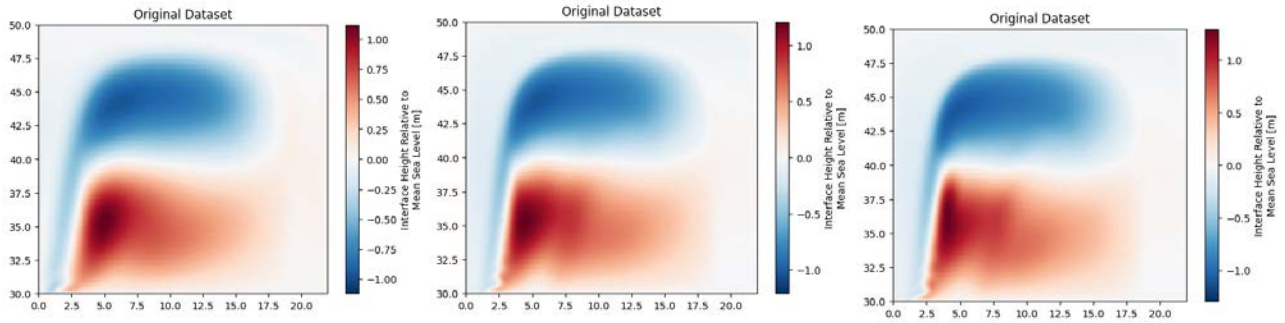




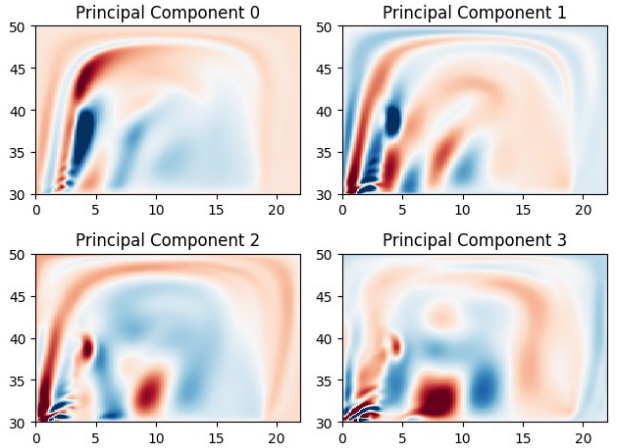


Intelligent sampling for multidimensional data

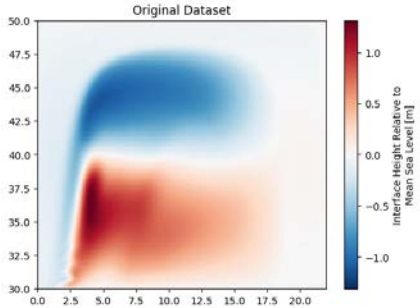
- Problem: Applying approach to images does not scale
  - One image is a single point represented by [NX x NY x NFeatures] -> High dimension!
- Solution:
  - Generate an initial dataset
  - Perform PCA to get principal components
  - Use metric of similarity between new snapshot and dataset basis [Mahalanobis Distance, ϕ]
  - Chi-squared (with given p-value) to determine whether to add new snapshot



PCA

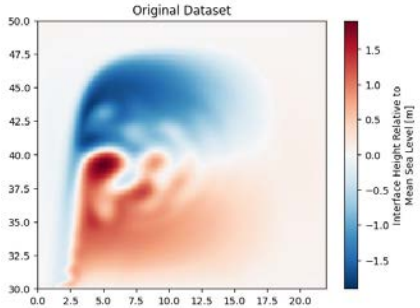


New Sample



Reject: $\phi < \chi^2$

New Sample

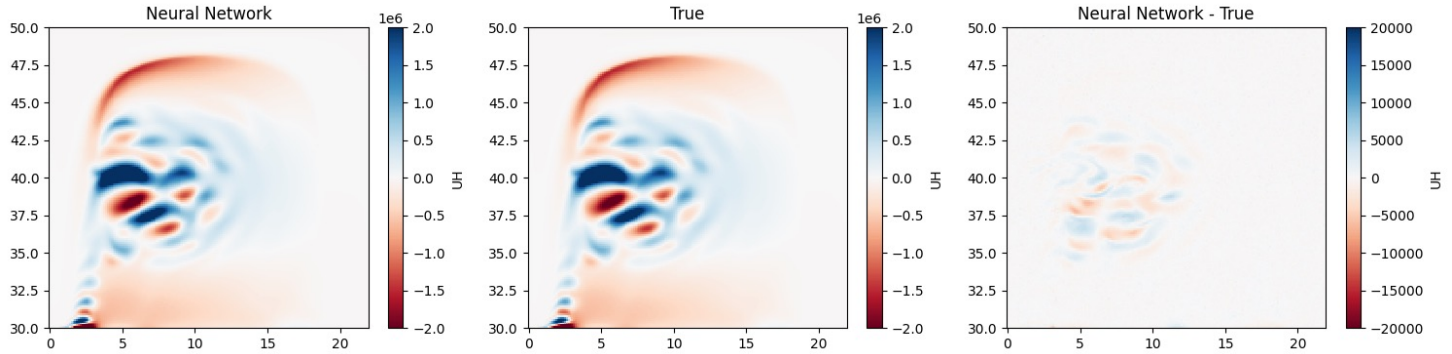


Accept: $\phi > \chi^2$

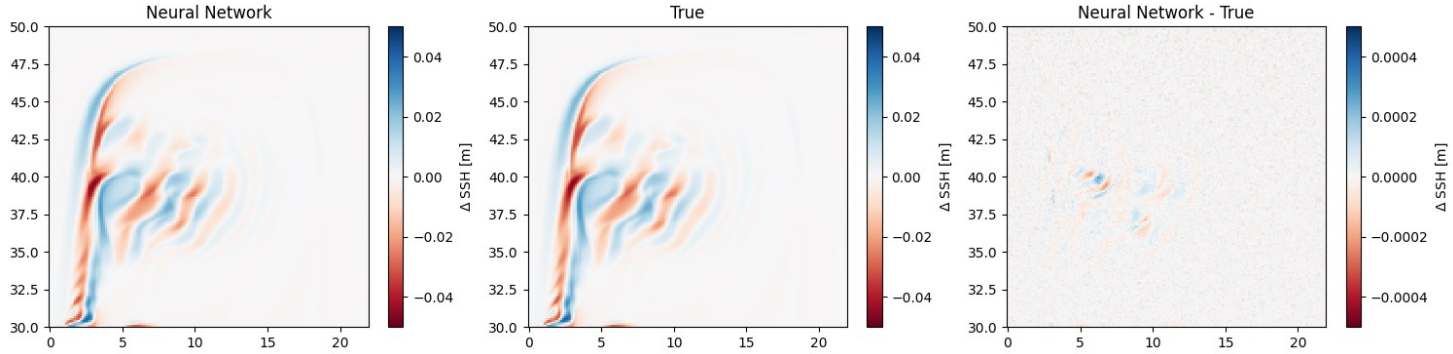
Neural network accurately predicts mass transports

- Accuracy:
  - Transport: ~1%
  - Column height: 1%
- Performance: ~20x increase
  - 1 A100 vs 128 CPUs
- GPU ‘unrolls’ timestepping loop
- Problem is almost linear
  - Potential additional savings with longer sequences
  - Tested up to 16

Zonal Transport [1st sequence]

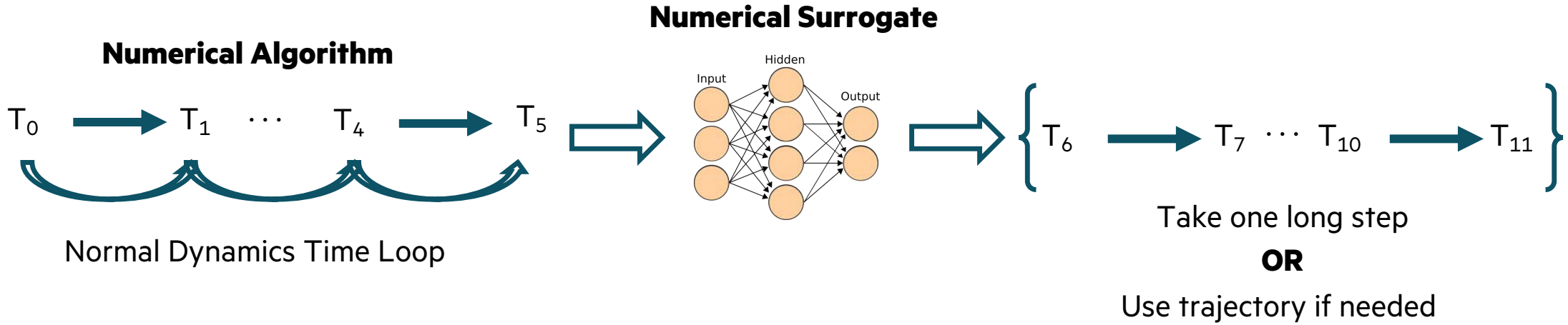


Change in column height [Over entire sequence]



A proposal for accelerating simulations

- **AI does not allow for formal numerical analysis**
  - Cannot estimate order of accuracy (truncation errors)
  - Cannot bound point-by-point errors
  - Errors likely accumulate
    - Weather surrogates dissipate
    - LLMs generate nonsense when trained on LLM-generated data
- **Numerical algorithms**
  - Formulations exist to dampen/control noise
  - Others simply need a pre-conditioner
- Proposed algorithm:
  Alternate AI surrogate and numerical method



Future Work

Accelerating the ocean dynamical core

- Explore other model architectures to find a more compact/simpler model
  - Build in rotational/reflective symmetries
- Embed inference from the AI surrogate into the dynamical core of MOM6
- Test generalization to different resolutions/cases
  - Should be invariant (just like the original numerical algorithm)
- Use online training to continually “fine-tune” model for the next sequence of predictions

Online, continuous learning

- Framework for training most types of surrogates sidesteps problems with training pipelines
  - Applicable to full surrogate, subgrid-scale, and numerical surrogates
- Intelligent sampling [not shown] can allow for simpler AI models especially for non-uniformly distributed systems

Testing at scale

- Intelligent sampling/online training framework has been tested at scale for CFD codes
  - Surrogate of particle-particle turbulence from multiphase solver
  - Effectively trained on ~47TB of data from 33 ensemble members
  - Single GPU for training was sufficient
- Redis database is a challenge for realistic ocean simulations distributed over many processors
  - New solution being developed with early release around October

Questions?

Andrew.Shao@hpe.com

For more information about SmartSim: https://craylabs.org