Augmenting Covariance Operators with Machine Learning: Generating Dedicated Datasets in the Cloud and a Prototype Model

Sergey Frolov[1]

Timothy A. Smith[1,2], Peter Vaillancourt[1,2], Jeffrey Whitaker[1], Zofia Stanley[1,2], Wei Huang[1,2], Henry R. Winterbottom[3], Clara Draper[1]

[1]NOAA Physical Sciences Laboratory (PSL)

[2]Cooperative Institute for Research in Environmental Sciences (CIRES), CU Boulder

[3]Lynker Technologies/NOAA/EMC/EIB

UIFCW 2023
A UFS Collaboration Powered by EPIC

Responding to disruptive Machine Learning technologies for NWP

Sergey Frolov will take the blame for controversial and provocative statements

Timothy A. Smith[1,2], Peter Vaillancourt[1,2], Jeffrey Whitaker[1], Zofia Stanley[1,2], Wei Huang[1,2], Henry R. Winterbottom[3], Clara Draper[1]

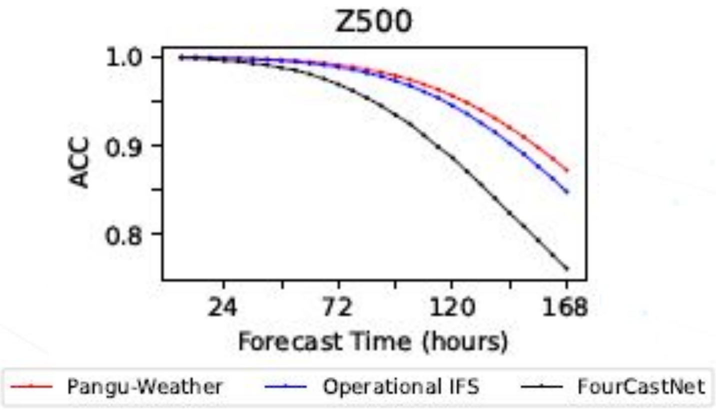
NOAA Physical Sciences Laboratory (PSL)

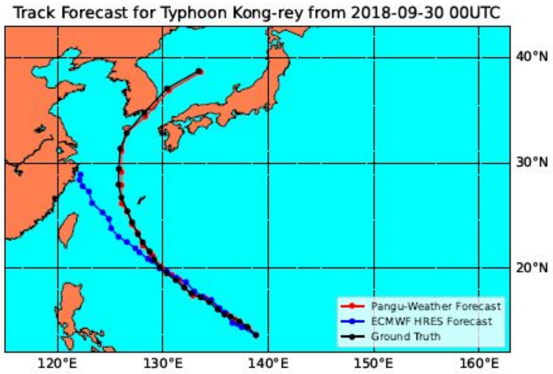
[2]Cooperative Institute for Research in Environmental Sciences (CIRES), CU Boulder

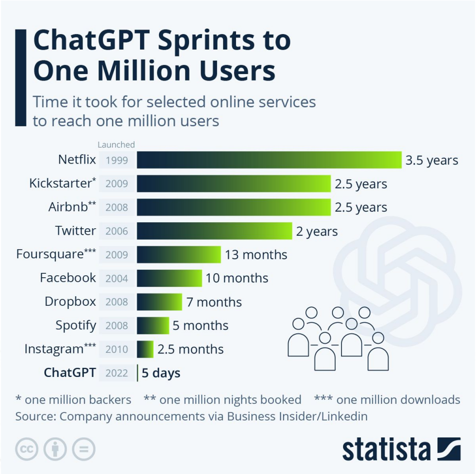
[3]Lynker Technologies/NOAA/EMC/EIB

UIFCW 2023
A UFS Collaboration Powered by EPIC

Machine learning: an existential threat to the NWP model?





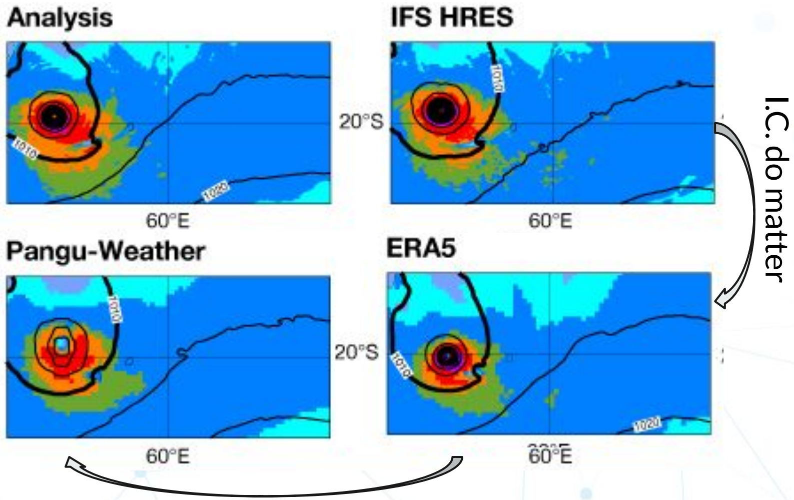


- Over the last 18 month, ML models (trained on ERA5) demonstrated performance competitive to ECMWF HRES forecast (ECMWF 2023)
- **ECMWF treat ML as an existential threat and a transformative opportunity to their business-as-usual model. And so should NOAA!!!**

UIFCW 2023
A UFS Collaboration Powered by EPIC

https://arxiv.org/pdf/2211.02556.pdf

A more nuanced look



Current generation of ML is too diffusive and may lack physical structure

ML models are trained on a 10-year old ERA5 technology yet are competitive with the state-of-the art:

- Initial conditions from the operational state-of-the-art model do matter! Unique role for NOAA operations.
- Current generation of ML models is too diffuse and possibly not dynamically consistent? This is getting improved by the external community.
- Current generation of ML models was not designed for data assimilation. A niche for NOAA research.
- High-quality training datasets are of paramount importance. A new-ish opportunity for NOAA.

UIFCW 2023
A UFS Collaboration Powered by EPIC

https://arxiv.org/pdf/2211.02556.pdf

NOAA PSL perspective and focus

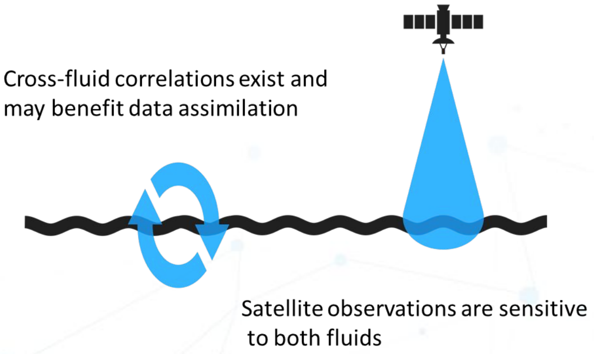
- Focus on producing high-value, cloud-ready, ML-ready training datasets:
    - 1957-present replay of the UFS coupled model to high-fidelity external analysis (ERA5/ORAS5);
    - Native coupled reanalysis and reforecast with UFS
    - Short hero runs with extremely large ensemble counts (upto 800 members)

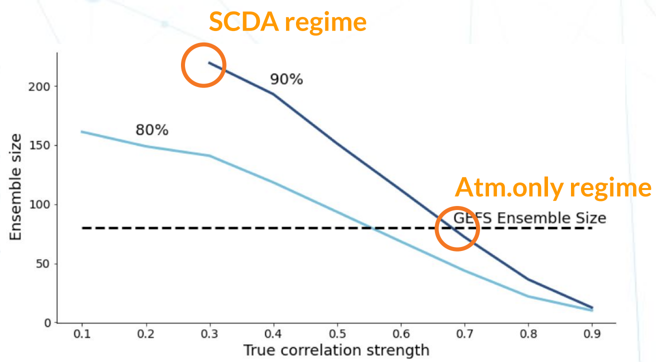
- Cloud-ready, ML-ready perspective:
    - Use NODD to allow users to co-locate NOAA datasets with computation
    - Move away from legacy output formats (grib, netcdf, flat files) to cloud-ready formats (zarr, netcdf+kerchunk)
    - Provide data on grids suitable for ML development

- Focus on ML development for data assimilation:
    - Operator replacement in DA
    - Perturbation models for ensemble propagation

UIFCW 2023
A UFS Collaboration Powered by EPIC

Case study: Enabling Strongly Coupled DA





Strongly coupled data assimilation:

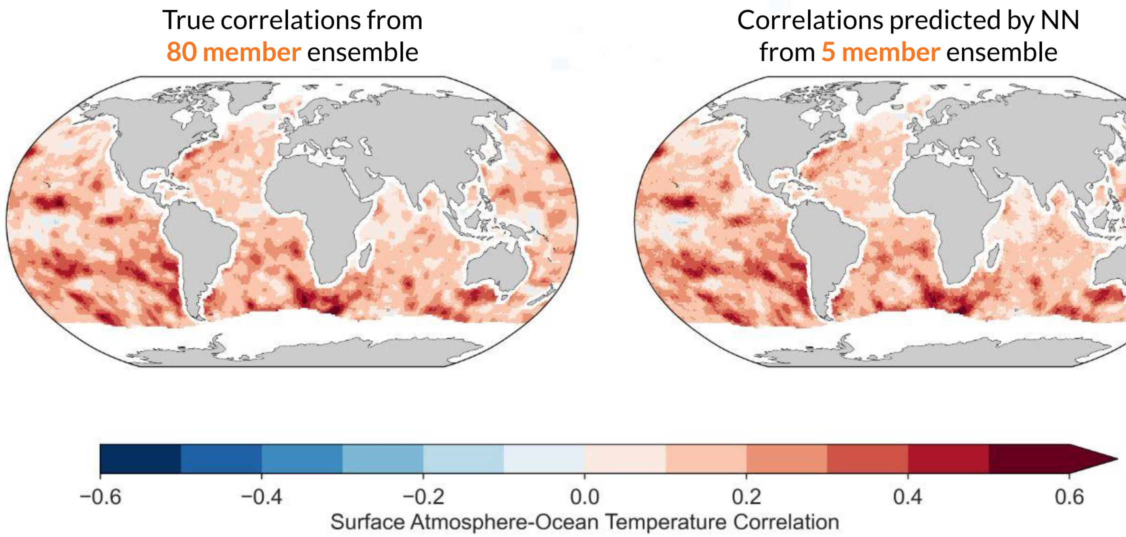
- Allows observations from atmosphere to impact ocean, and vice versa;
- Expected improvement in S2S & hurricane forecasting.

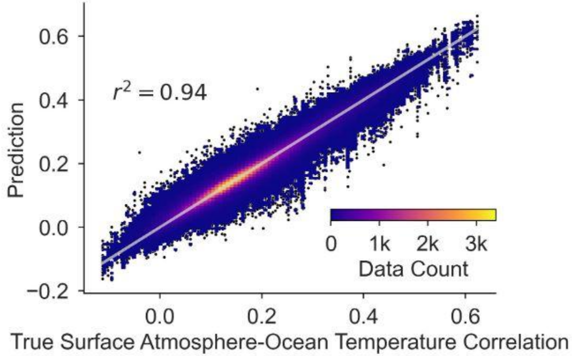
However, cross-domain covariances are **intermittent** and **low amplitude**

Many more ensemble members are required to accurately estimate covariances

How to reduce this cost in order to enable SCDA?

Prototype: Predicting AST-SST Correlation





The correlation structure from the 80 member ensemble is captured well by 5 members + neural network

See github.com/NOAA-PSL/mlcdc for details

Current work: Dedicated Datasets for ML+DA

Expand original prototype

- 1 degree, 800 members, spanning 3 months
- ¼ degree, 240 members, spanning 1 month
- Data will be generated using RDHPCS cloud allocation
- Resulting datasets will be made publicly available in Zarr format through NODD

Challenges:

- **It is extremely hard to support this work using competitive NOAA funding**

HIGH TECHNOLOGY FOUNDATION
BUILDING A STRONGER WEST VIRGINIA

UIFCW 2023
A UFS Collaboration Powered by EPIC

Conclusions

- The last 18 months of groundbreaking results from the ML community challenge our existing NWP business model.

- NOAA has a role to play in the emerging need and opportunity for:
  - Producing training data for ML models
  - Co-locating NOAA data with computational opportunities in cloud-native, ML-native formats
  - Investing in ML research to augment and transform our current operational stack.

END: Questions

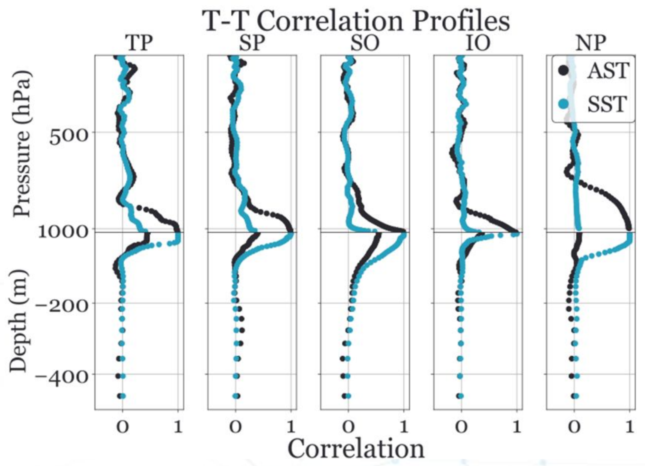
Constructing a Neural Network Vertical Correlation Model

Training, validation, & testing dataset

- Weakly coupled atmosphere & ocean UFS model
- 80 members
- Single 24-hr forecast

Architecture

- Feed forward neural network
- Input: 5-member average surface quantities (e.g., 2m temperature & humidity, SST, mixed layer depth)
- Output: vertical temperature correlation, as if we had used 80 members
- Each grid cell is treated independently

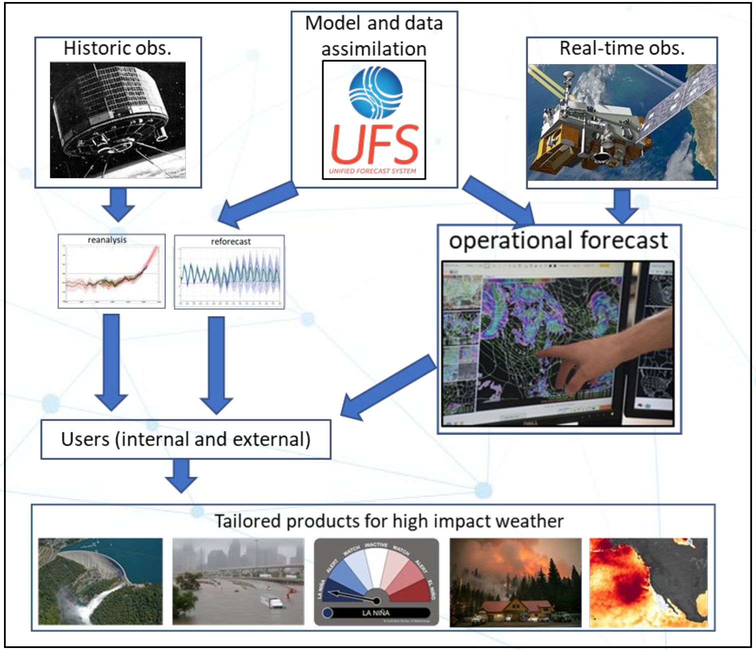


Atmosphere-ocean correlations from ~1,000 member ensemble

Main question: is the correlation signal predictable, based on a very small ensemble average of surface quantities?

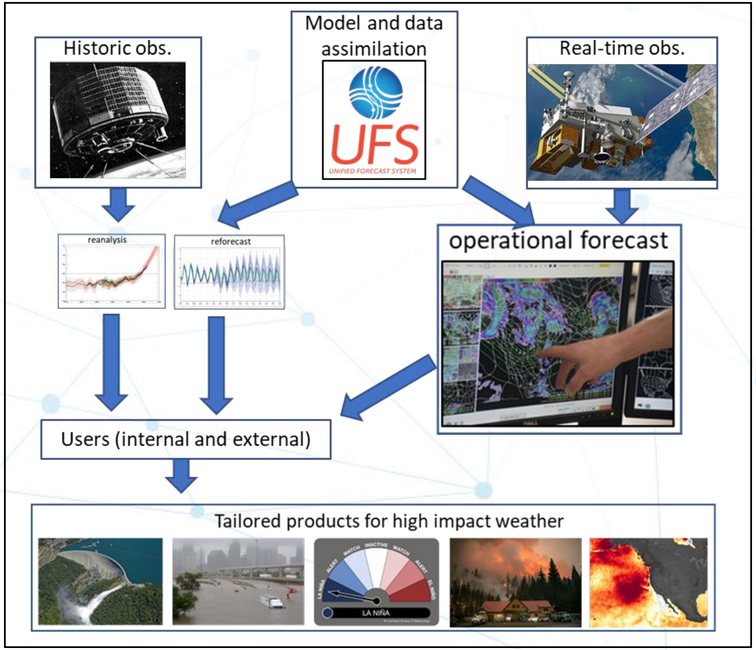
How can NWP enterprise adjust to the ML era?

Existing NOAA/NWP model:
focused on forecasts

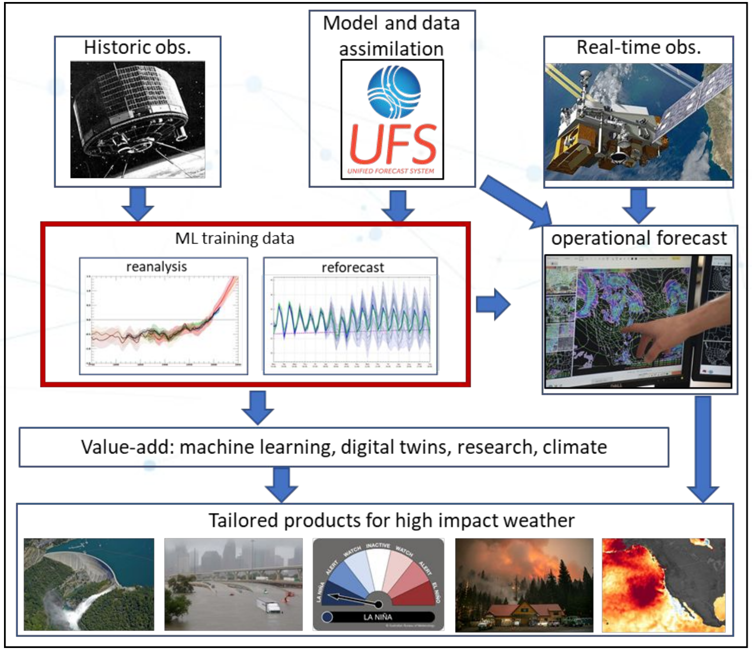


How can NWP enterprise adjust to the ML era?

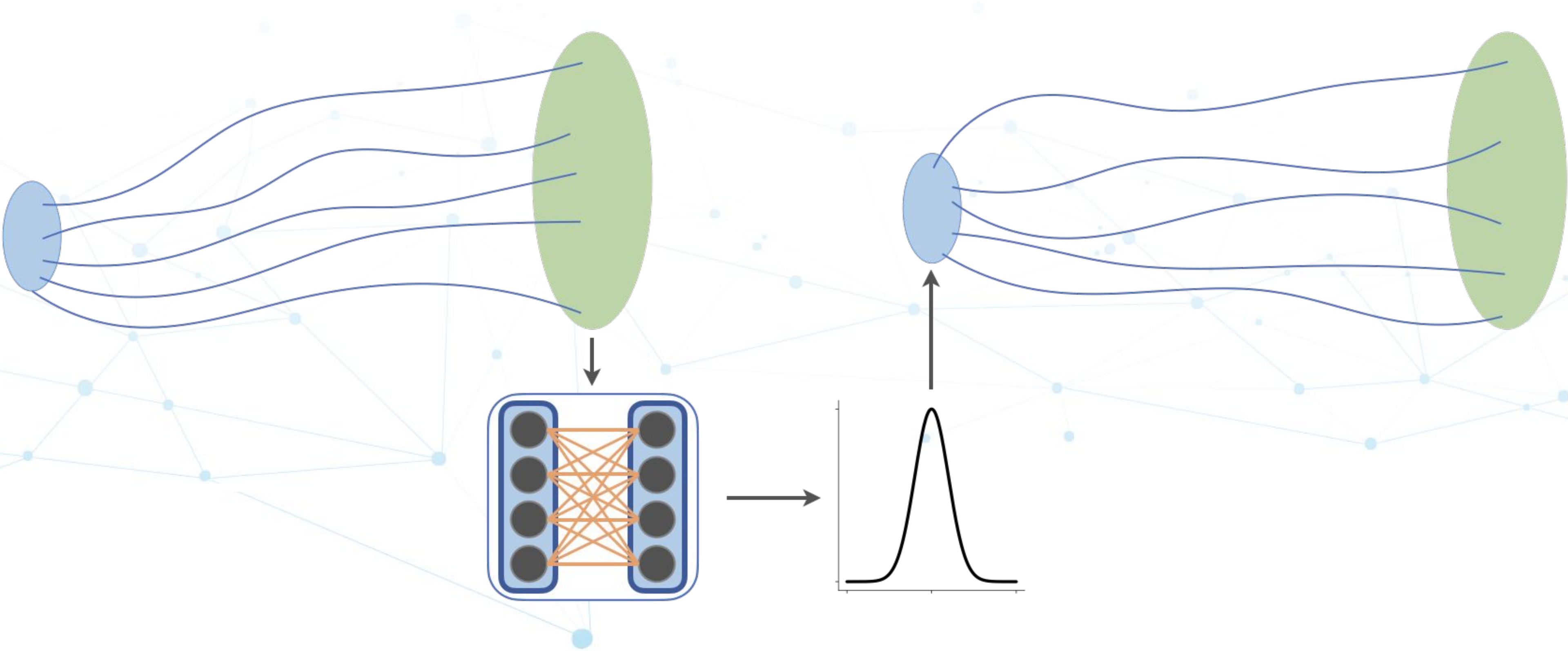
Existing NOAA/NWP model:
focused on forecasts



New model?
ML training datasets are equally important to external users and quality of operational forecasts



Proposed concept: augment a small ensemble with a neural network



Status quo: operational ensembles require many members and parameterizations