Reconstructing Line P: A virtual hackathon comparing data-driven methods for imputation

Hayley V. Dosser, Emily O'Grady, Susan E. Allen, James R. Christian and **Amber M. Holdsworth**





Fisheries and Oceans Canada | Pêches et Océans Canada
Environment and Climate Change Canada | Environnement et Changement climatique Canada



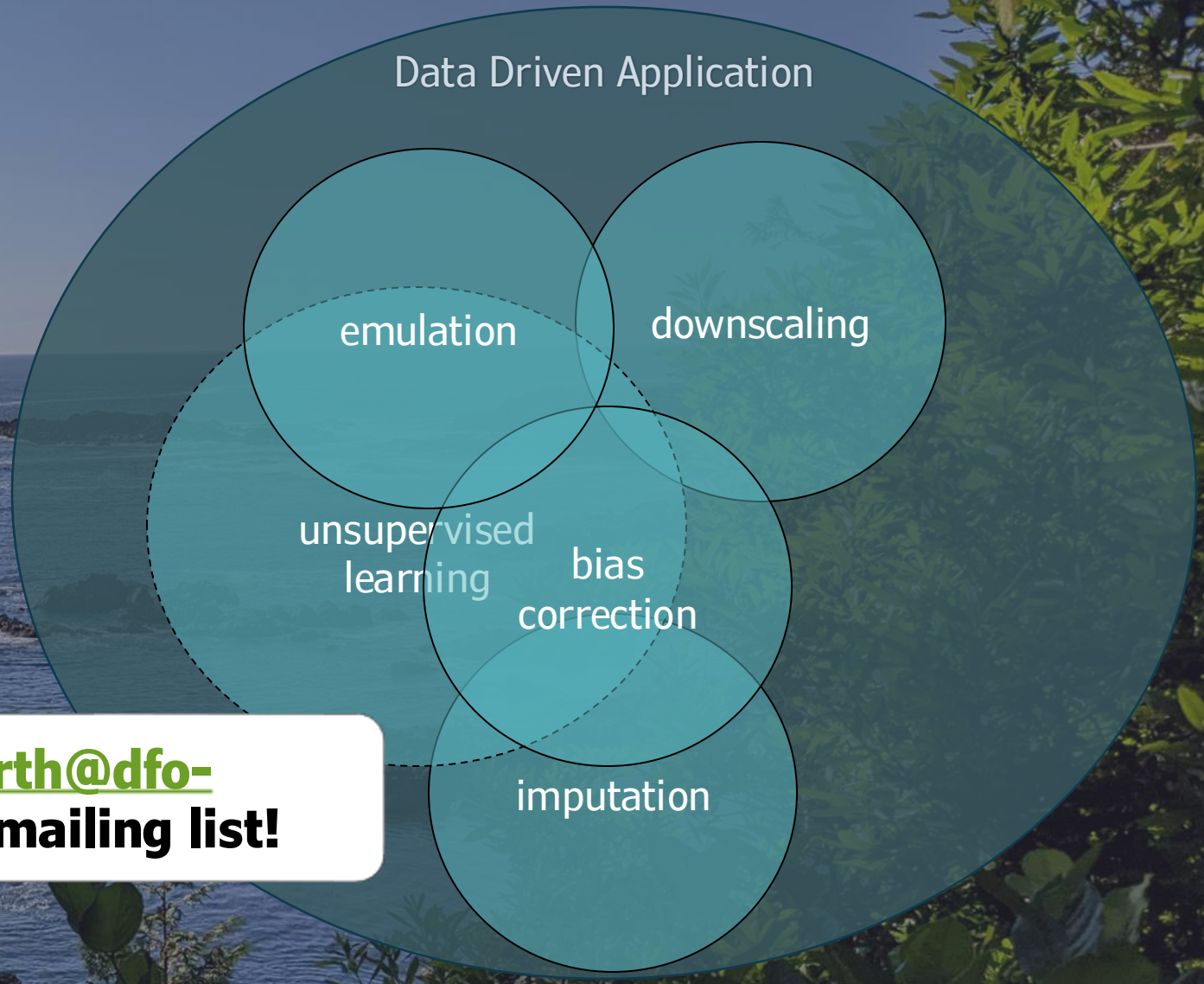
THE UNIVERSITY OF BRITISH COLUMBIA

Machine Learning Methods for Ocean Data (ML-Ocean) Working Group

Who: International expertise from oceanography, atmospheric science, statistics and ecology (government, academia, and ngos)

What: Community, Seminars and Discussion Sessions and Hackathons.

Why? ML advances rapidly and advancement requires an understanding of the oceanographic data.

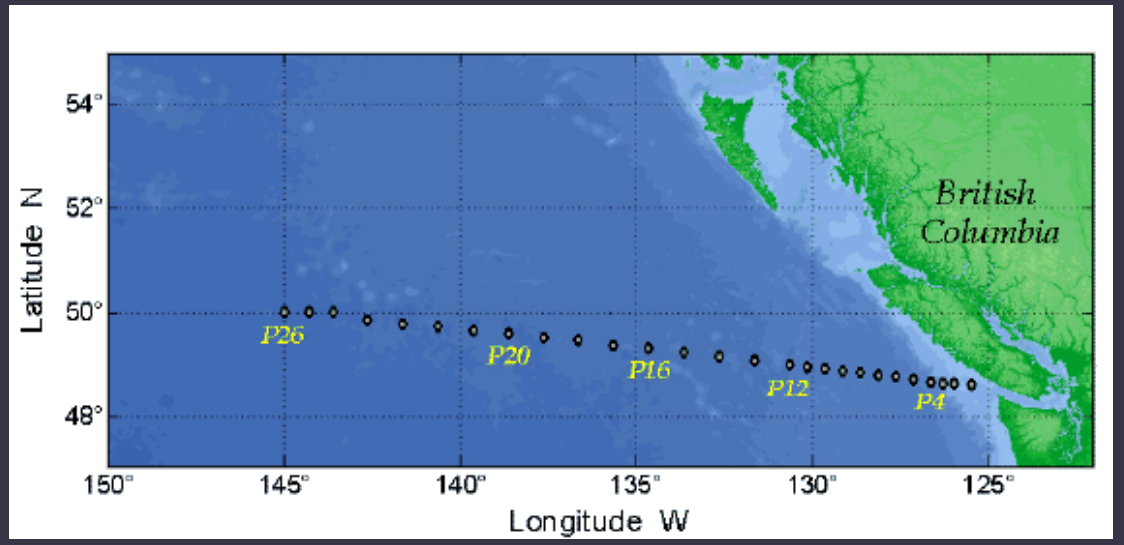


Email Amber.Holdsworth@dfo-mpo.gc.ca to join the mailing list!

Fisheries and Oceans Canada
Pêches et Océans Canada

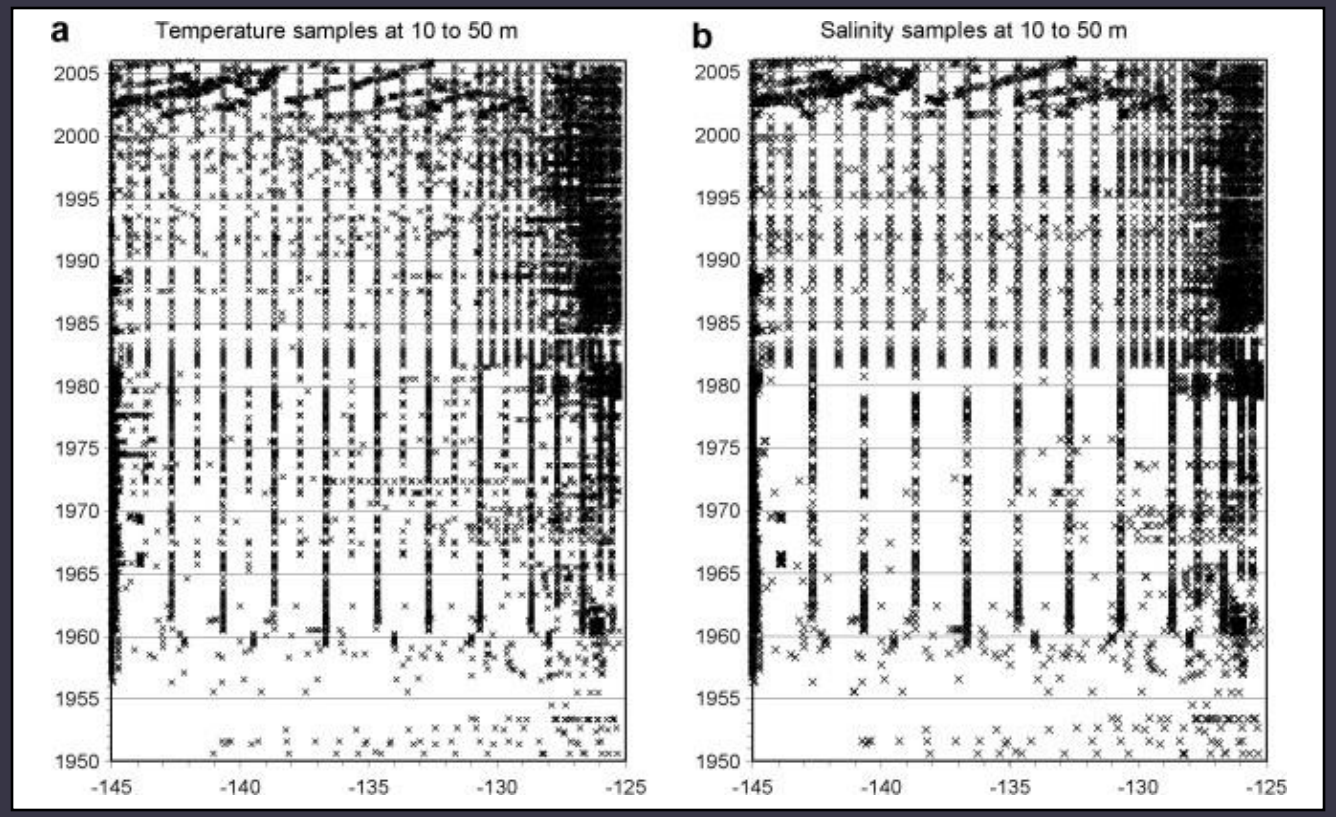
Line P hydrographic transect

- 27 stations covering 1425 km
- Oceanographic sampling from 1949 to now
- Biogeochemical samples were used from 1990-2019
- Multiple DFO cruises per year
- Vertical profiles to 2000 m depth



What is “imputation”?

Imputation:
the process of estimating and replacing missing data

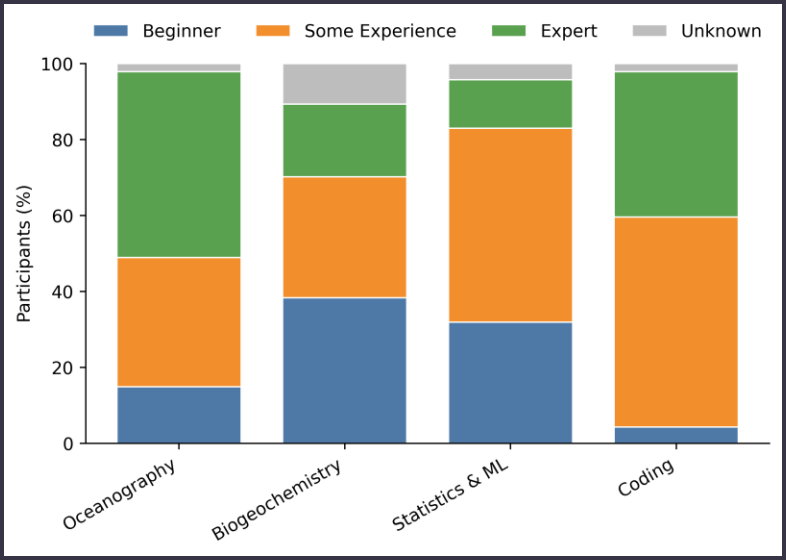


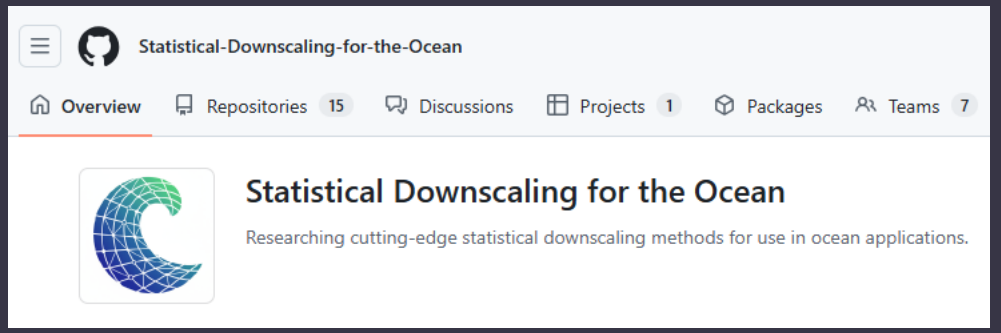
Temporal and spatial distribution of temperature and salinity samples along Line P, including some Argo data from 2001 onwards. *From Crawford et al. 2007*

Reconstructing Line P - Event

Hackathon Participation:

- 2-week, virtual event
- Seven teams, 4-6 participants per team, range of experience levels, 7 countries
- Various career levels from students to late career, almost half were early career





Hackathon Tools:

Participant guides:

- Complete Hackathon guide
- Training Data guide
- Evaluation guide

Data and codes:

- Template codes on GitHub
- Training data hosted on the GoC GPSC
- Command-line verification tool

Other tools: Slack, Google Drive, Zoom

Hackathon goals

Primary goal

- Starting from the DFO Line P hydrographic datasets, generate a complete dataset (no missing values) for monthly temperature (T), salinity (S), dissolved oxygen (O_2), dissolved inorganic carbon (DIC), and total alkalinity (TA).

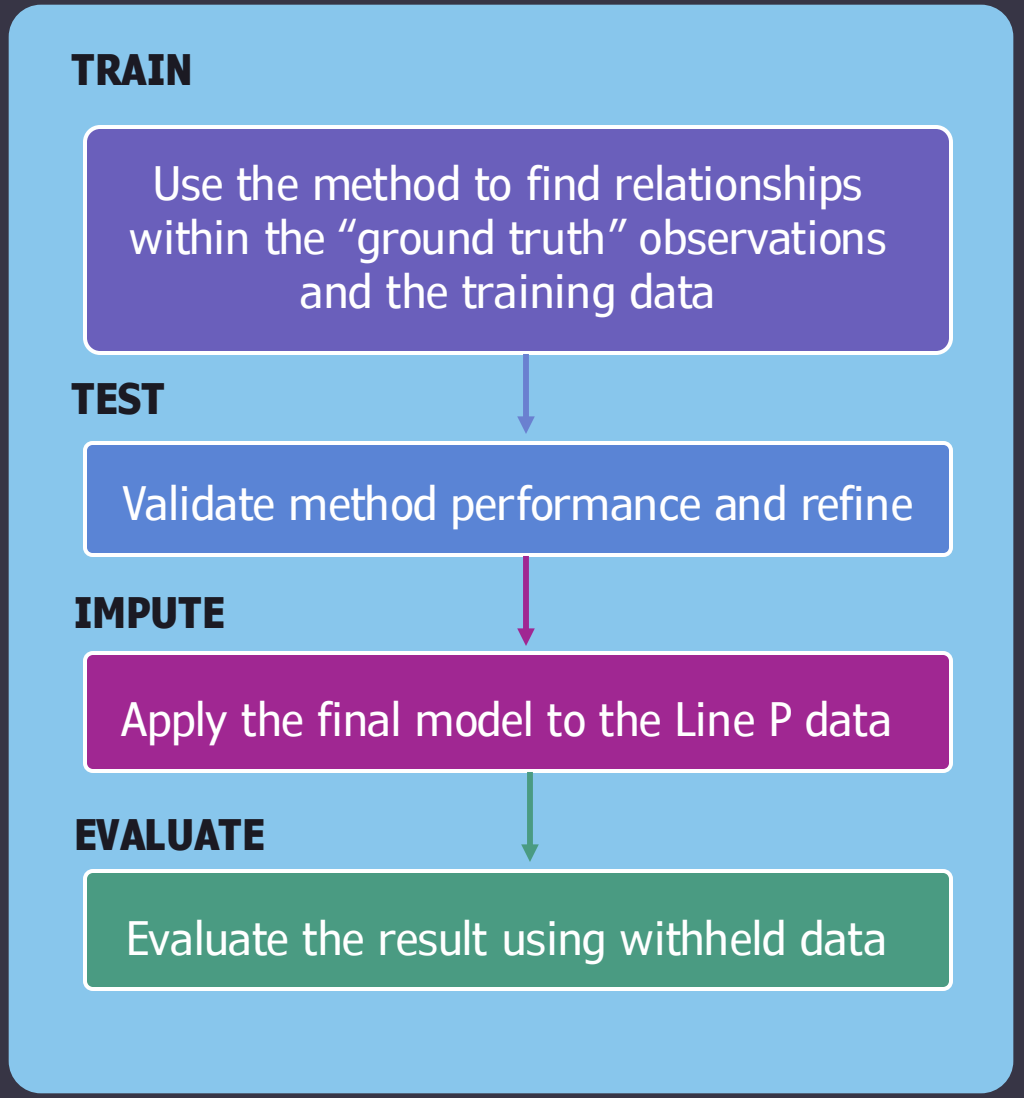
Stretch goal

- Attempt to reconstruct the fields between the Line P stations.

Reconstructing Line P - Methods

Data-driven methods for imputation:

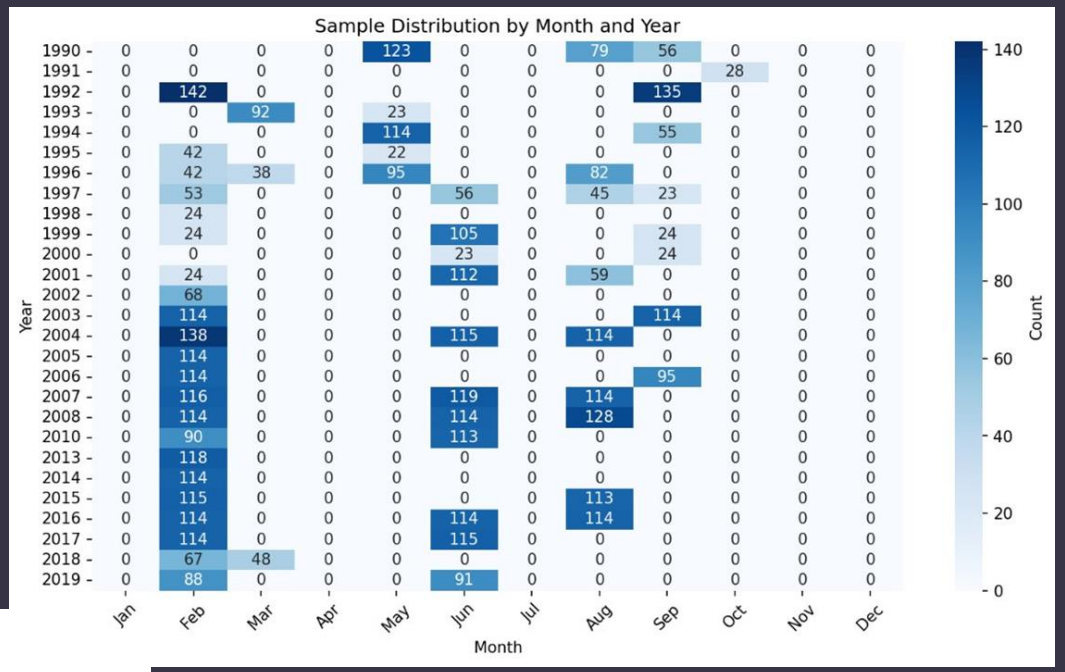
Method	Type
Neural Networks	Machine learning
Fourier Neural Operators	Machine learning
Graph Neural Networks	Machine learning
Gaussian Process Regression	Machine learning
Quantile Mapping	Statistical
Transfer Learning	Machine learning

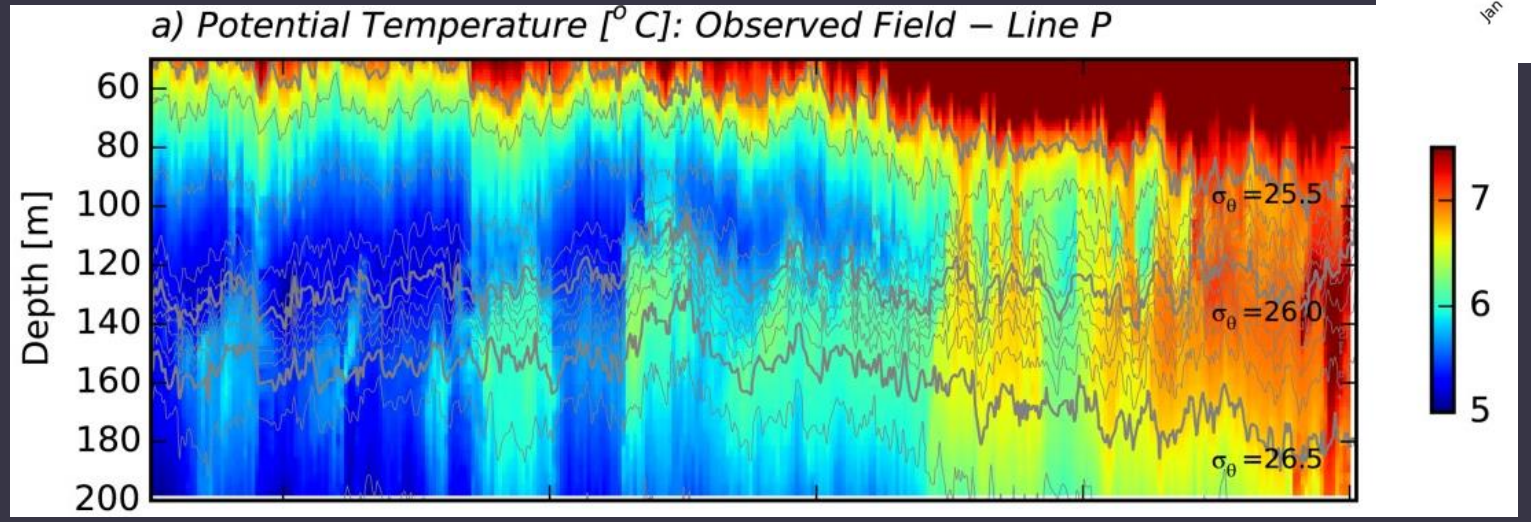


The Hackathon **was not** a rigorous methodological intercomparison. Instead, it brought together oceanographers, statisticians, and machine-learning practitioners **to explore** how different data-driven approaches can handle oceanographic data.

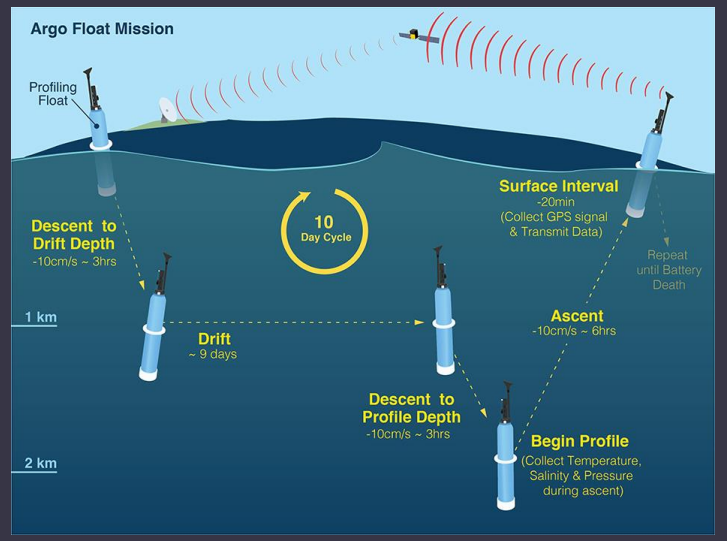
Challenges in an oceanographic context

1. Sparse and biased sampling (stations, seasons, decades, variables)
2. Variability across scales
3. Non-linear relationships between variables
4. Strong gradients and heterogeneous environments
5. Climate trends superimposed on natural variability (e.g., PDO, NPGO, ENSO)





Example showing many scales of variability along Line P from observations (in Klymak et al., 2015)

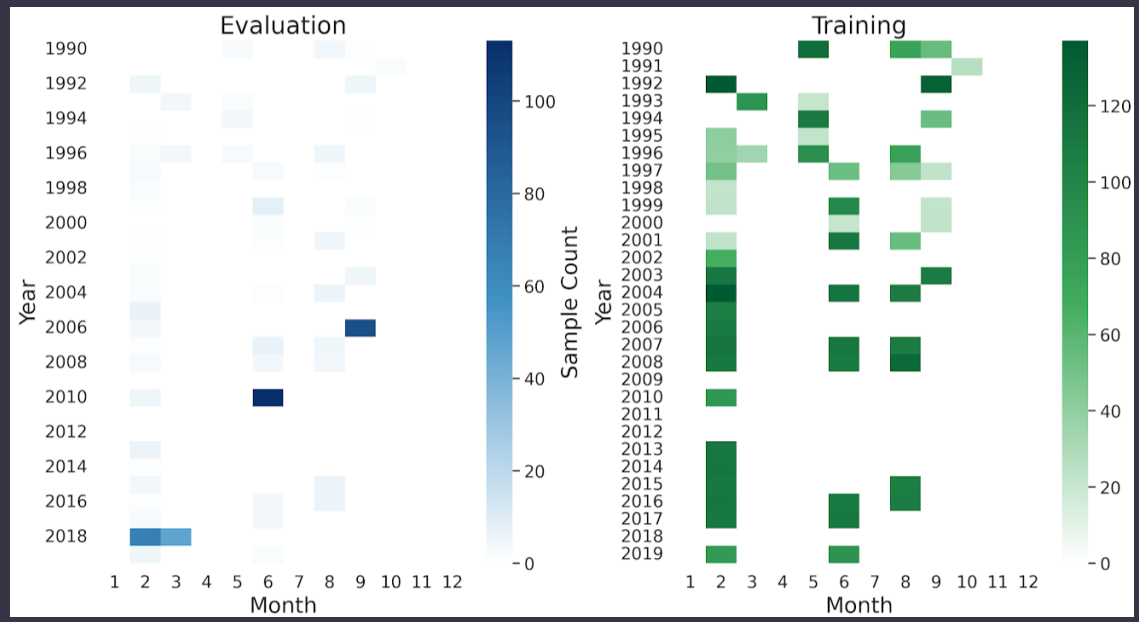


Hackathon data was carefully curated

CTDs



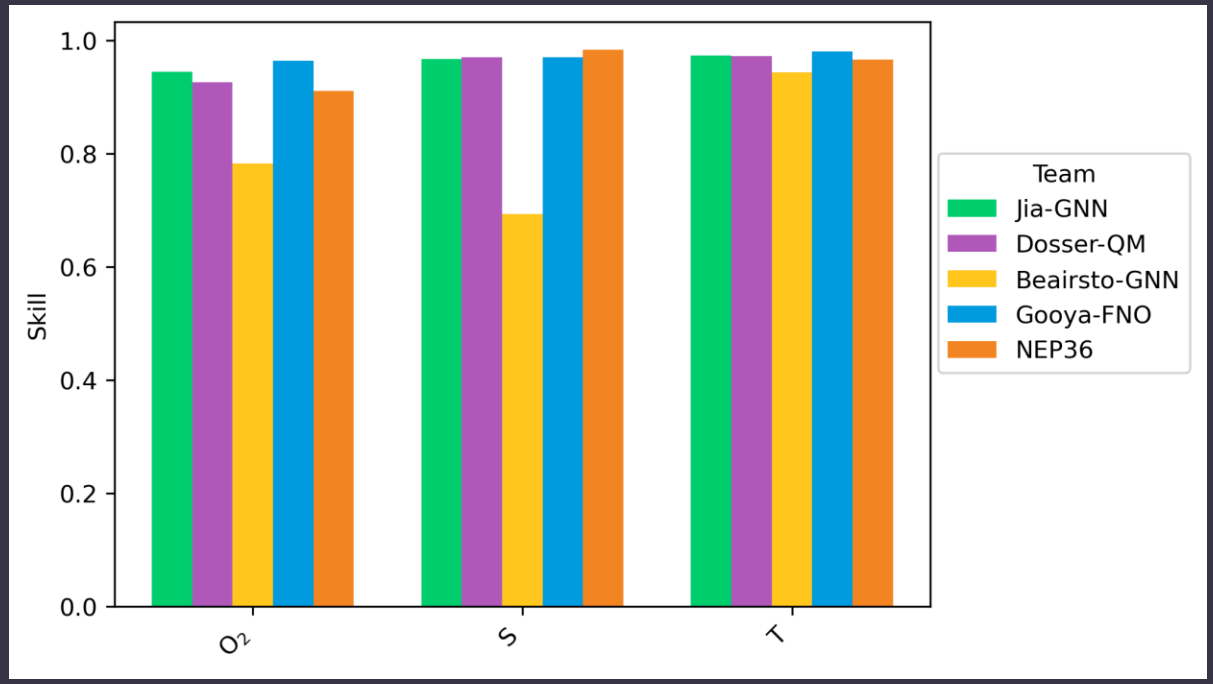
Bottles



1. Each team was restricted to a curated set of training data
    1. High resolution models and Global Models
    2. QA/QC'd Line-P CTD and Bottle data
    3. Argo data
2. Some of the Line P observations were separated from the training data and withheld for evaluation

Evaluation and the end of the event

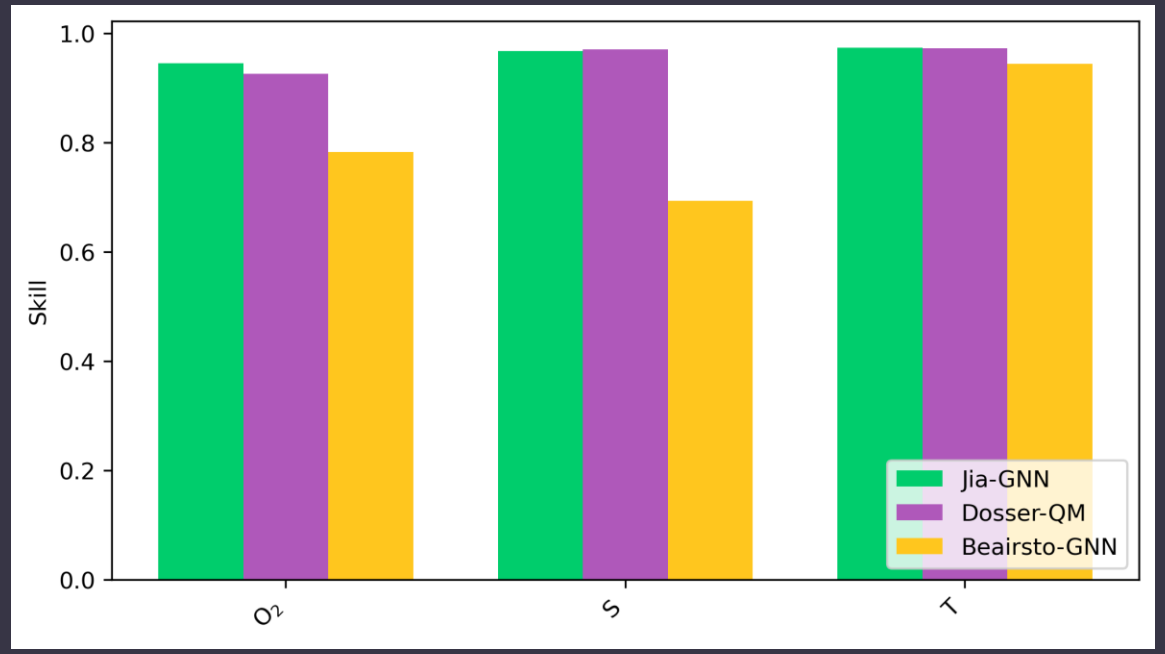
1. Hackathon participants presented their methods and final results during the closing ceremony
2. The coordinator shared the final scores (RMSE, correlation, outlier penalties)
3. The winners were announced and granted the title of “Line P Hackathon Hero”



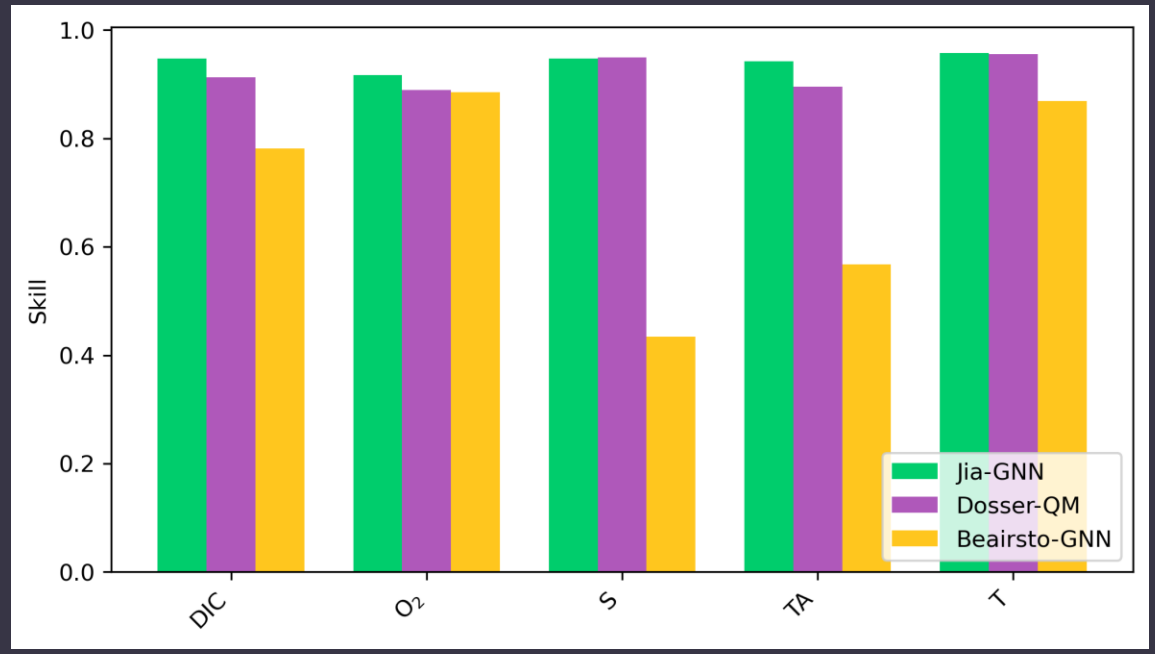
Evaluation for a subset of T and S over 1996-2020



Evaluation and the end of the event



Evaluation for CTD profile variables



Evaluation for bottle samples including BGC

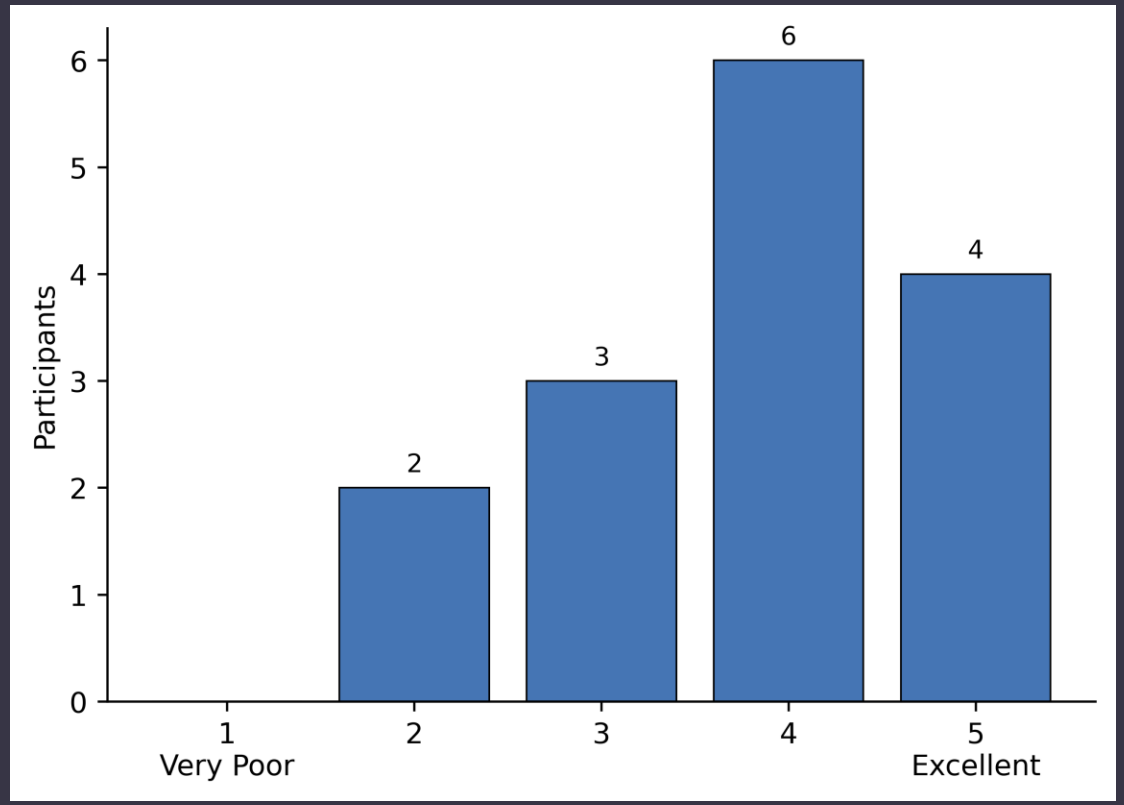
Conclusions and Next steps

Many data-driven approaches hold promise for imputation along Line P

A collaborative network emerged from the event

- Technical Report
- Rigorous evaluations of reconstruction methods using a perfect model framework
- Generation of open-source data products and code-bases

Overall experience ratings from the exit survey.



We'd like to hear from you!

Would you use a reconstructed line P dataset?

What for? What is most important to evaluate for your application?