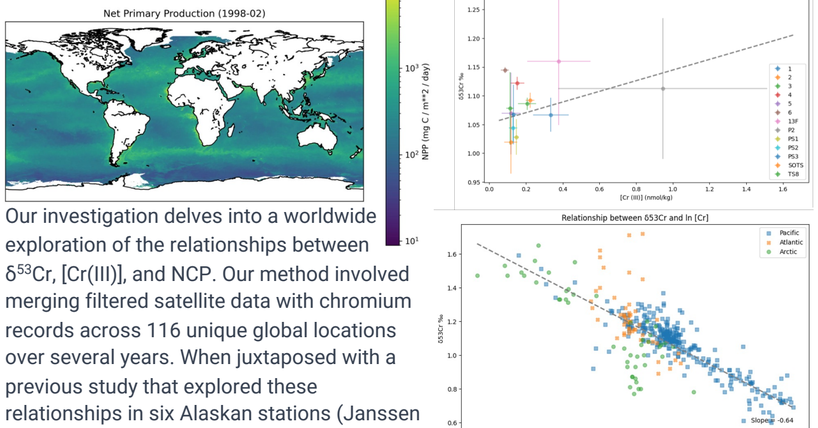
Understanding the Ocean Biological Carbon Pump with Big Data

Team Members: Echo Chen, Tianshu Feng | Project Lead: Zuchuan Li, Nicolas Cassar | Project Manager: Shrikant Chand

Background

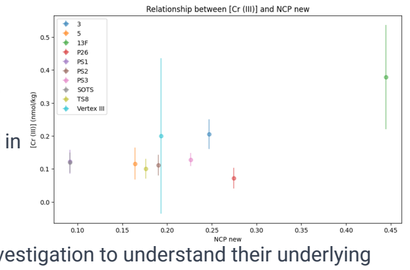
The ocean, absorbing about 80% of Earth's heat and 30% of our carbon dioxide, is crucial for climate regulation. A key factor in understanding this process is **Net Community Production (NCP)**. NCP directly correlates to the ocean's carbon saturation as it calculates the difference between respiratory consumption and photosynthetic production.This project aims to use satellite data and cruise records to uncover relationships between oceanic measurements and build an accurate model for predicting NCP. Our goal is to better understand how the ocean affects

Exploration



Our investigation delves into a worldwide exploration of the relationships between δ^{53}Cr, [Cr(III)], and NCP. Our method involved merging filtered satellite data with chromium records across 116 unique global locations over several years. When juxtaposed with a previous study that explored these relationships in six Alaskan stations (Janssen et al., 2020), our larger dataset revealed intriguing results.

The data reveals a strong linear relationship between [Cr(III)] and NCP, similar to the localized findings from the Alaskan stations. The correlation between δ^{53}Cr and [Cr(III)], however, displayed an opposing linear trend in our global study. This could be due to the broader range of [Cr(III)] in our data and our station grouping methods.

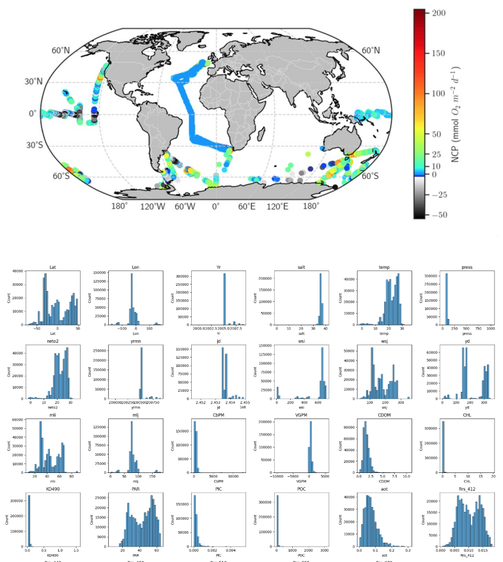


These intriguing disparities call for further investigation to understand their underlying reasons. See the visualized data for a closer look at these relationships.

$\delta^{53}Cr_{,,}$: Chromium isotope ratio signaling redox conditions.

[Cr(III)]: Concentration of trivalent Chromium, indicating geochemical processes.

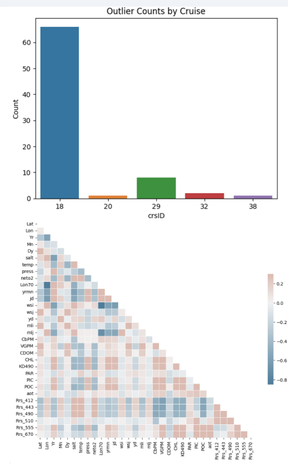
Data Description



Our dataset comprises 355,296 records spanning from 1999 to 2008, with an extensive range of 38 parameters. These include fundamental parameters such as time, location, and temperature, alongside satellite-collected parameters like VGPM (a type of NPP measurement) and CHL (the concentration of chlorophyll a).

The accompanying map illustrates the global distribution of NCP values, providing a holistic picture of the ocean's spatial carbon balance. The 24 key feature distribution plots offer an in-depth look into individual data characteristics.

Data Preprocessing

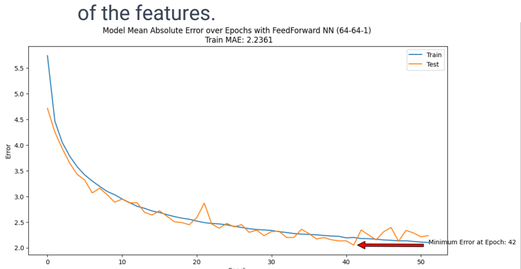


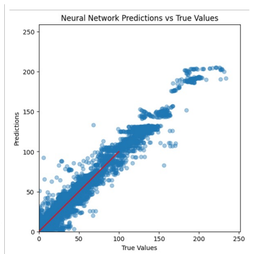
Our streamlined preprocessing workflow, outlined in the above flowchart, ensures data quality and reliability for accurate analysis. We began with cleaning, addressing duplicates, missing data, and outliers. Next was feature engineering to enrich our model, and feature selection for optimal relevance. Finally, we split the data for performance evaluation. This meticulous process is crucial to extract meaningful insights from every piece of data, driving a comprehensive understanding of oceanic carbon balance.

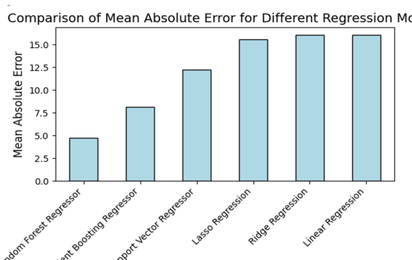
Results

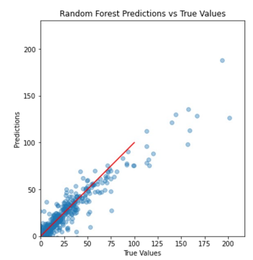
Our analysis employed two primary models for evaluation, each revealing distinct insights into the data.

- FeedForward Neural Network (64-64-1): The training of our neural network spanned over 100 epochs, with a notable Mean Absolute Error (MAE) of 2.1998 at the 42nd epoch. A comparison of Neural Network Predictions versus True Values demonstrates the model's predictive capabilities.
- Ensemble & SVM Models: In our comparison of Decision Tree, Random Forest, and Support Vector models, the Random Forest model outperformed the others, achieving a training MAE of 1.7677 and a testing MAE of 4.7631. The model's Feature Importance plot and the Predictions versus True Values graph provide further understanding of its performance and the significance of the features.







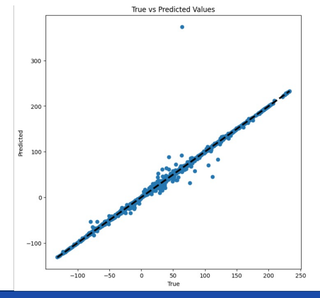


Discussion and Conclusion

Aiming to strike a balance between interpretability, accuracy, and computational complexity, our choice of models and metrics were carefully considered. The Feedforward Neural Network offers the potential for high accuracy in predictions, while the Random Forest model provides an interpretable structure. Mean Absolute Error (MAE) served as our chosen metric due to its simplicity and effectiveness in quantifying prediction accuracy.

A crucial aspect of our study was addressing the challenges of spatial autocorrelation in our data. Traditional random sampling methods for train/test splits are not appropriate for our spatial data as they might yield overly optimistic model performance estimates. To circumvent this, we employed a stratified train/test split, creating spatial bins based on latitude instead of performing random splits. This approach was validated when a random split yielded an implausible 100% correct model, underscoring the importance of a spatially-conscious approach in our study.

Our findings demonstrate the potential of machine learning models in predicting NCP, contributing to a better understanding of the ocean's role in the global carbon balance, and highlighting the importance of nuanced data handling in spatial studies.



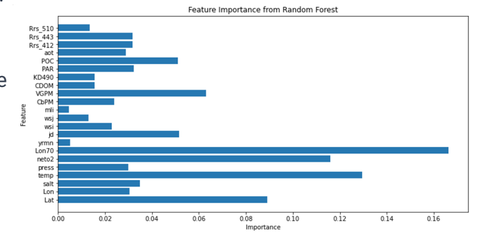
Future Work

While our initial models have provided valuable insights, there are several promising directions for further study:

- Advanced Spatial Models: Spatial lag models and kriging could potentially offer more nuanced insights into spatial autocorrelation in our data.
- Performance Evaluation & Model Selection: Testing models on unseen data will provide a robust performance assessment. There is a need to further balance accuracy, interpretability, and computational efficiency.

Finally, leveraging the insights gained from the Random Forest's feature importance plot, we plan to explore the predictive relationship between NCP and the features more closely. Modifying our feature set could potentially enhance the performance and interpretability of our models.

Our journey into oceanic carbon balance continues, and we believe our exploration will contribute significantly to climate science, providing crucial insights into the world's most extensive carbon sink - the ocean.



References

Janssen, D. J., Rickli, J., Quay, P. D., White, A. E., Nasemann, P., & Jaccard, S. L. (2020). Biological control of chromium redox and stable isotope composition in the Surface Ocean. *Global Biogeochemical Cycles, 34*(1). https://doi.org/10.1029/2019gb006397

Li, Z., & Cassar, N. (2016). Satellite estimates of net community production based on O_2/Ar observations and comparison to other estimates. *Global Biogeochemical Cycles, 30*(5), 735–752. https://doi.org/10.1002/2015gb005314

To explore our GitHub repository, simply scan the QR code in the corner. We encourage open collaboration and discussion to advance this important climate science study.