AI for Life Science - 2

Identification of exogenous variables for forecasting GRACE time series data

Xinyue GAO (xinyue.gao2000@outlook.com)
Azeez LIADI (Liadiazeez3@gmail.com)
Omar LAHAM (p.omarlahham@gmail.com)

Introduction

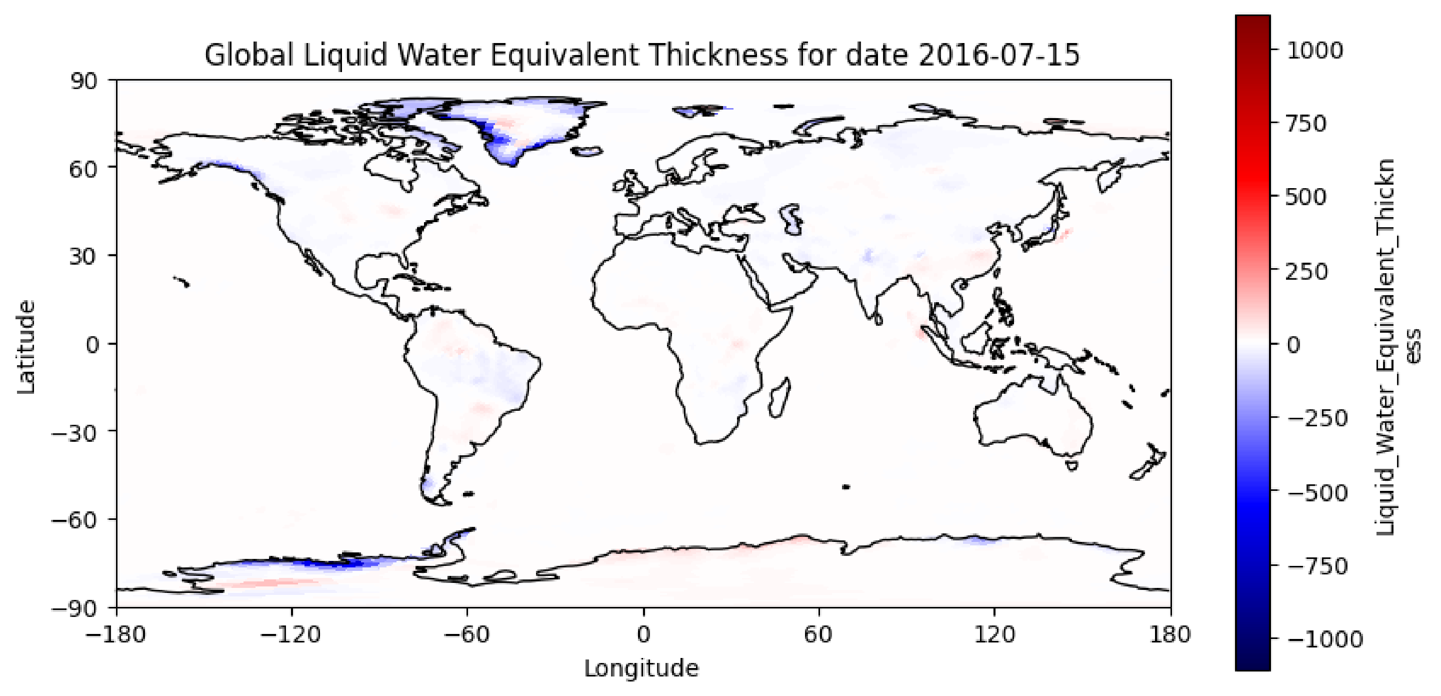
1 GRACE Data Overview

Gravity Recovery and Climate Experiment (GRACE)

A satellite mission launched in 2002

Tiny changes in Earth's gravitational field caused by mass movements such as water movement, ice melt, and **groundwater depletion**

Global map in time series, from 2002-04-18 to 2024-04-16



2 Project Goal

Forecast GRACE data in future

Sustainable water management, agricultural planning, and environmental conservation

Increasing water demands and climate variability

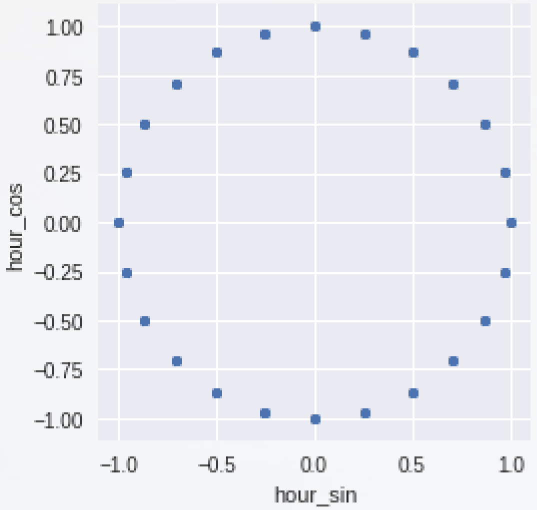
3 Key Aspects

Forecast groundwater levels using machine learning models trained on GRACE data and exogenous variables.

Choose 6 exogeneous variables to forecast GRACE ground water time series.

Find the most contributing variable.

Method - Approach of preprocessing on data



1 Dividing into patches

Divide the longitude (360°) * latitude (180°) into 20°*20° grids as a unit to study.

In total: 162 grids after the division.

2 Averaging variables on cosine-weighted

A cosine-weighted average of a variable over time, with adjustments for the spherical nature of the Earth.

Accurate representation of regions near the poles (appear smaller in flat projections, more significant in real-world calculations)

3 Cleaning data

Set all dates into the 1st of the month. A **linear method** is used to replace the missing values.

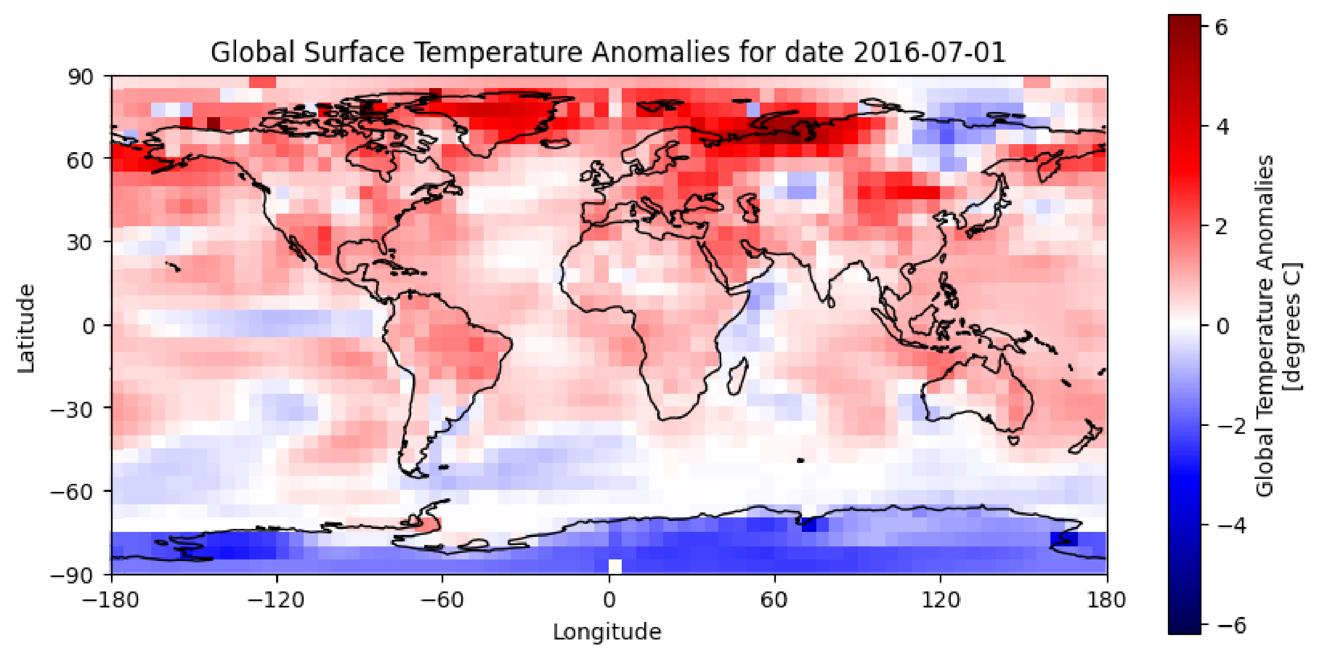
4 Encoding cyclical features

Let deep learning know that features such as months occur in cycles. It gives us date_cos and date_sin as a result.

	sea_level	temperature	precipitation	relative_humidity	clwc	liquid_water_thickness	eddy_kinetic_energy	month	year	quarter	month_sin	month_cos	quarter_sin	quarter_cos
2002-04-01	-0.007504	0.341324	6.086234	82.212124	0.000003	0.633437	513.02	4	2002	2	0.866025	5.00e-01	1.22e-16	-1.00e+00
2002-05-01	0.034153	0.367918	4.980250	80.664751	0.000004	1.064181	633.27	5	2002	2	0.965926	2.58e-01	1.22e-16	-1.00e+00
2002-06-01	0.020669	0.284284	5.307215	80.178718	0.000003	-0.905211	552.54	6	2002	2	1.000000	6.12e-17	1.22e-16	-1.00e+00
2002-07-01	0.004759	0.317057	3.011250	79.029472	0.000002	-2.874602	686.65	7	2002	3	0.965926	-2.58e-01	-1.00e+00	-1.83e-16
2002-08-01	-0.010802	0.066045	3.754239	76.767973	0.000001	-4.843994	673.21	8	2002	3	0.866025	-5.00e-01	-1.00e+00	-1.83e-16

Method - Exogenous Variables in Our Model - 1

Sea Surface Temperature



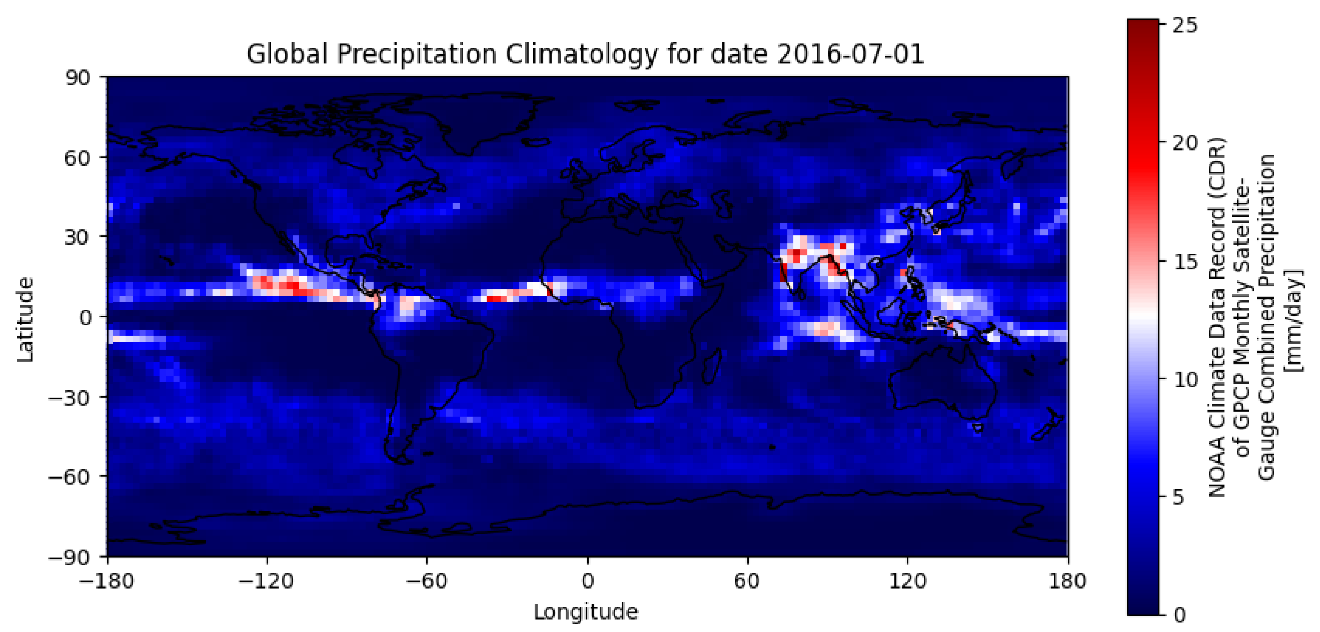
The temperature of the water's surface layer, typically measured in the top few meters of the ocean.

- Evaporation impact
- Snow and ice melt

From 1850-01-01 to 2023-12-01

Huang, Boyin, Chunying Liu, Viva Banzon, Eric Freeman, Garrett Graham, Bill Hankins, Tom Smith, and Huai-Min Zhang. "Improvements of the Daily Optimum Interpolation Sea Surface Temperature (DOISST) Version 2.1", *Journal of Climate* 34, 8 (2021): 2923-2939, doi: https://doi.org/10.1175/JCLI-D-20-0166.1

Precipitation



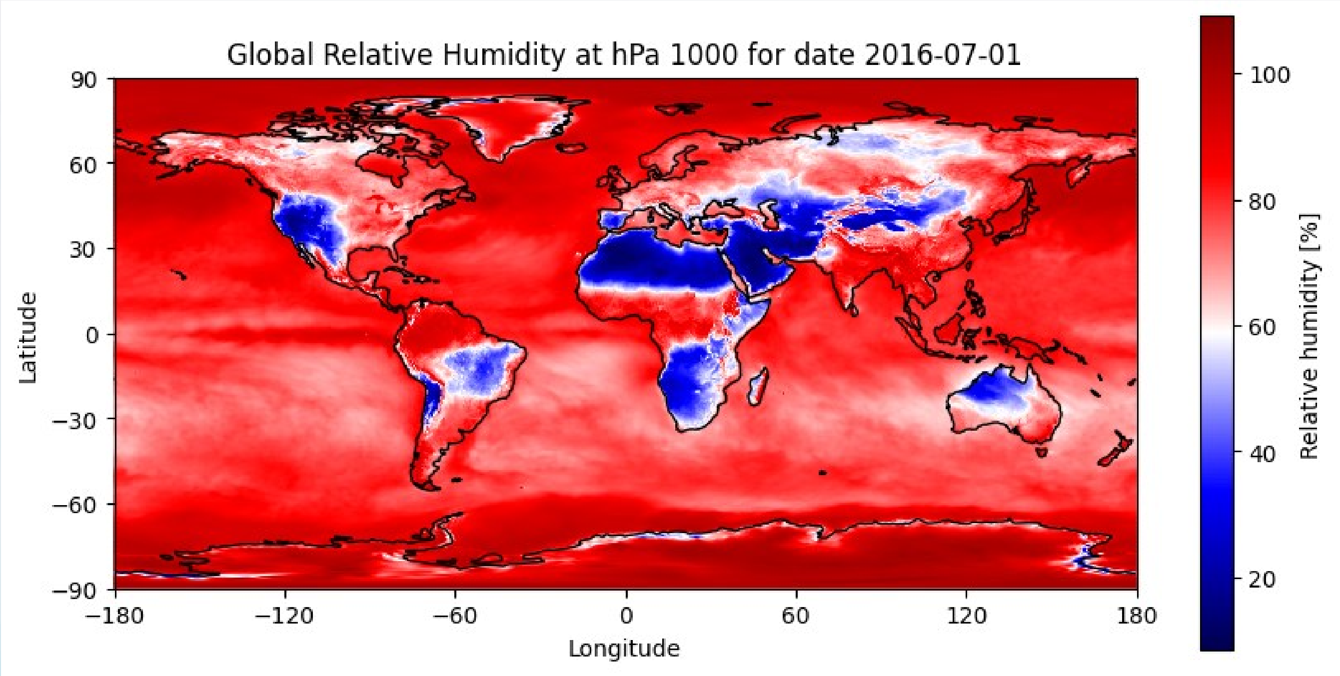
Water released from clouds in the form of rain, snow, sleet, or hail that falls to the ground.

- Primary source of recharge
- Regional variability

Adler, Robert F., Mathew R. P. Sapiano, George J. Huffman, Jian-Jian Wang, Guojun Gu, David Bolvin, Long Chiu, Udo Schneider, Andreas Becker, Eric Nelkin, and et al. 2018. "The Global Precipitation Climatology Project (GPCP) Monthly Analysis (New Version 2.3) and a Review of 2017 Global Precipitation" Atmosphere 9, no. 4: 138. https://doi.org/10.3390/atmos9040138

Method - Exogenous Variables in Our Model - 2

Relative Humidity



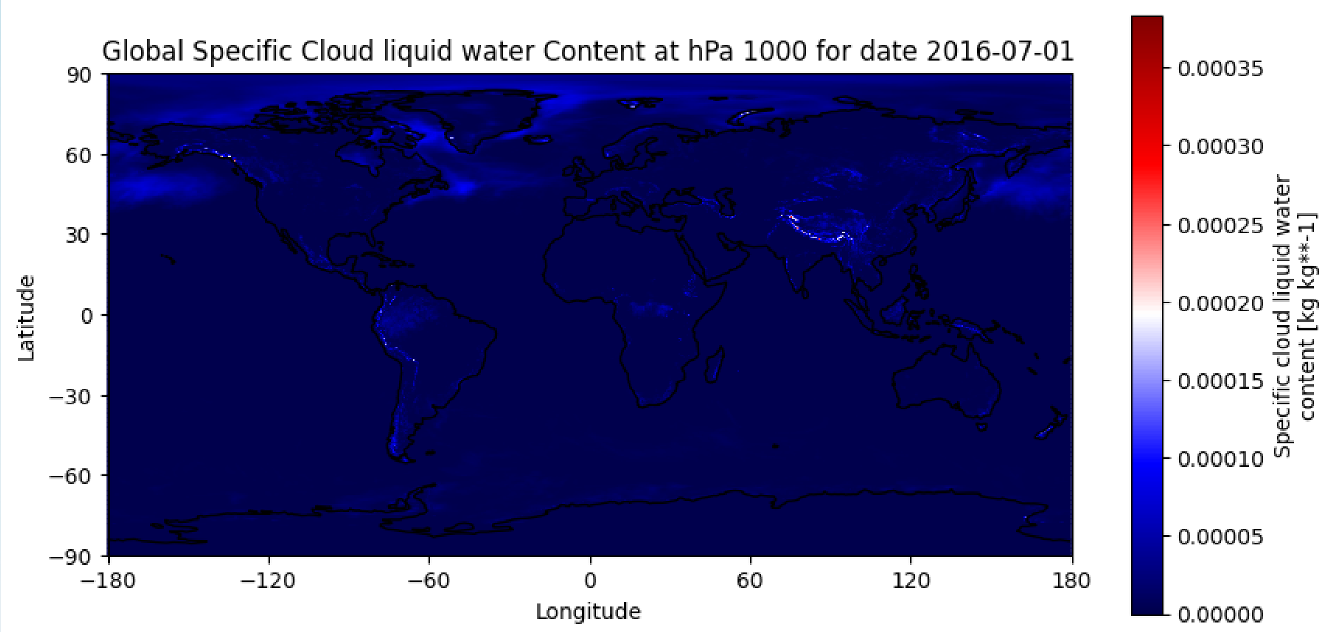
The percentage of water vapor in the air relative to the maximum amount of water vapor the air can hold at a given temperature.

- Precipitation Likelihood

From 2002-01-01 to 2024-08-01

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., Thépaut, J-N. (2023): ERA5 monthly averaged data on pressure levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), DOI: 10.24381/cds.6860a573 (Accessed on 09-09-2024)

Cloud Liquid Water Content



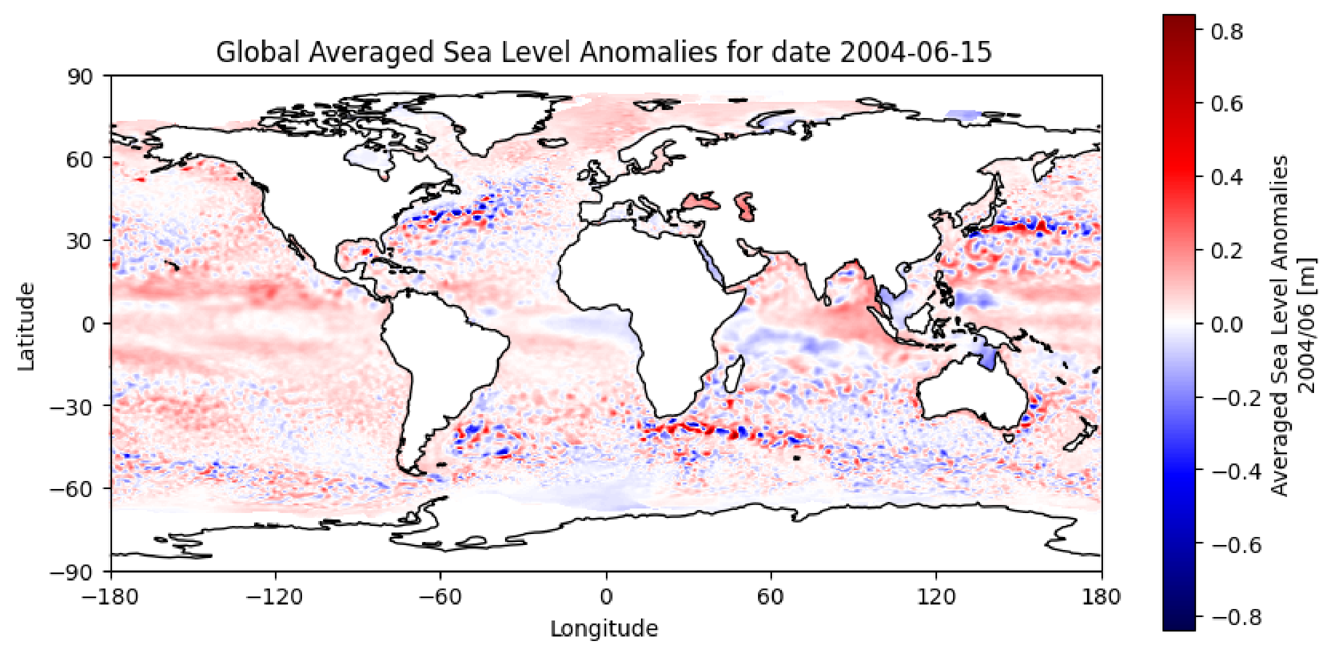
The mass of liquid water droplets contained in a cloud per unit volume of air, typically measured in grams per cubic meter.

- Indicator of Precipitation Potential
- Seasonal and Geographic Impact

From 2002-01-01 to 2024-08-01

Method - Exogenous Variables in Our Model - 3

Sea Level



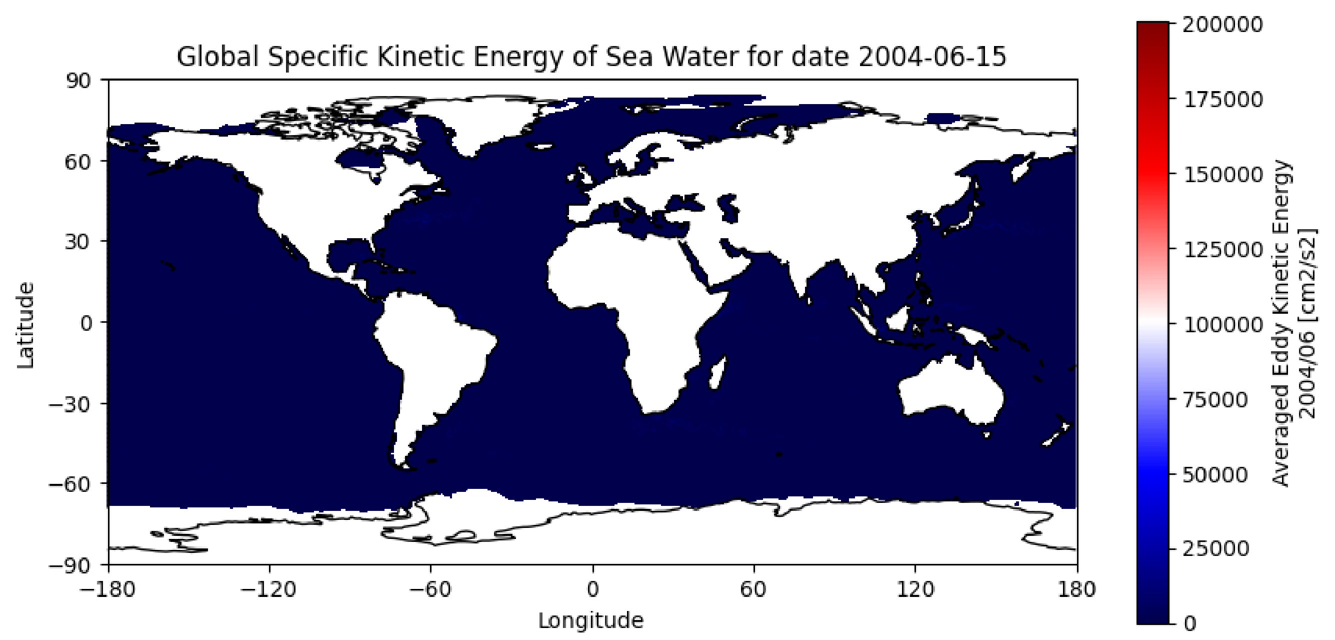
The average height of the ocean's surface, used as a standard in reckoning land elevation and measuring climate change impacts.

- Reduced Groundwater Discharge
- Recharge Disruption

From 1993-01-01 to 2023-08-01

Copernicus Climate Change Service, Climate Data Store, (2018): Sea level gridded data from satellite observations for the global ocean from 1993 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). DOI: 10.24381/cds.4c328c78 (Accessed on 23-09-2024)

Eddy Kinetic Energy (Ocean Currents)



A measure of the energy in ocean currents due to eddies (small-scale, circular movements of water), which are created by wind, the Earth's rotation, and interactions between different water masses.

- Ocean-Atmosphere Interactions
- Climate Influence

From 1993-01-01 to 2023-08-01

Method – Machine Learning to forecast

1

Data preprocessing

Splitting of datasets

Train set: 2002-04 --- 2018-01 (n = 90)

Validation set: 2018-02 --- 2023-08 (n = 68)

Test set: 2023-09 --- 2028-08 (n = 60)

2

Hyperparameter Tuning

Bayesian Search

Mean Absolute Percentage Error (MAPE)

3

Model Training

ForecasterAutoregMultiVariate

LGBMRegressor

Transformation: Standard scaling

Root Mean Squared Scaled Error (RMSSE)

4

Prediction and Future Forecasting

Prediction on 5 years (60 steps)

Feature importance: both lagged variables and exogenous variables

5

Batch Processing for Multiple Grids

Loops through grids.

For every grid:

- Model tuned, trained, and used to predict
- Prediction and feature importances are stored
- RMSEE is aggregated to overall error for the batch

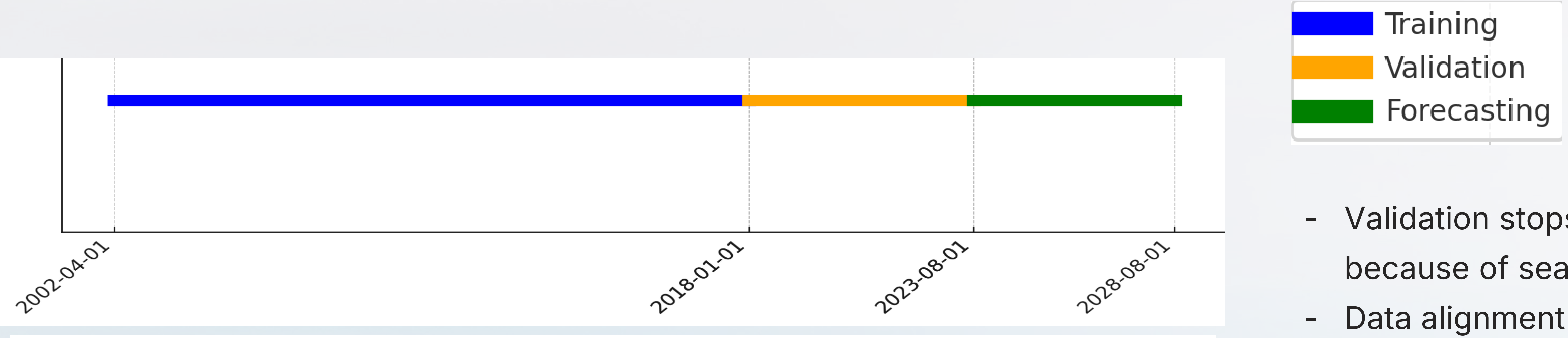
6

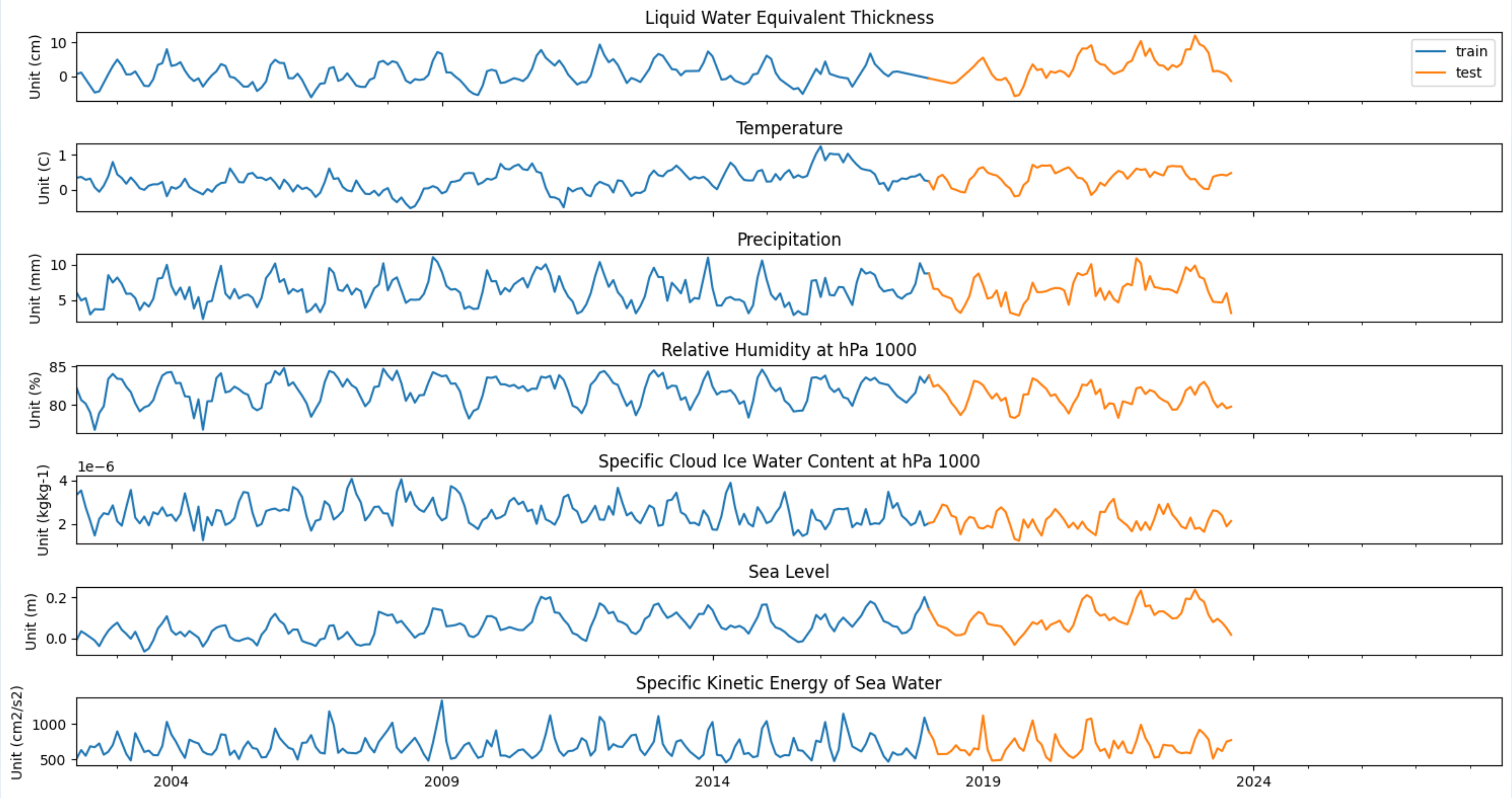
Final Output

After processing all datasets:

- A data frame of predicted values for each grid
- Overall RMSSE
- A list of important features that most contributed to the forecast

Method – Train and Validation





- Validation stops at 2023-08 because of sea level data
- Data alignment at this point
- Consistency across variables

	Feature	Importance
0	temperature_lag_8	397
1	sea_level_lag_1	348
2	eddy_kinetic_energy_lag_22	341
3	temperature_lag_20	330
4	relative_humidity_lag_13	324
5	relative_humidity_lag_22	300
6	liquid_water_thickness_lag_1	293
7	eddy_kinetic_energy_lag_20	283
8	relative_humidity_lag_1	283
9	clwc_lag_13	279

Results of importance

Overall RMSSE : 2.91483

Step 1 (2024-09-01):

Variable	Predictive power
Relative humidity	1892
Liquid water thickness	1742
Temperature	1585
Precipitation	1386
Sea Level	843
Cloud liquid water content	731
Eddy kinetic energy	502

Step 60 (2028-08-01):

Variable	Predictive power
Relative humidity	2189
Temperature	1726
Cloud liquid water content	1334
Precipitation	1048
Eddy kinetic energy	877
Sea level	838
Liquid water thickness	293

Conclusion

Implications and concerns

Different scale of data

- Unable to use the real GRACE data because of the missing in sea level
- Extrapolating based on older data between 2023-08 and 2024-04

Truncated Training data

- Unable to use the most recent data for target variable in training
- Potentially impact the performance, especially for near-term forecasts

Next steps

1. **Extend prediction on every step in 5 years**
   - Performance trajectory
   - Feature importance evolution
2. **Improvement of algorithms**
   - HistGradientBoostingRegressor
   - Also used in our task 1, gave better result than LGBM
3. **Enhanced Model Complexity**
   - More complex models (e.g. RNN, LSTM)
   - Also suitable for capturing long-term dependencies in time series

Thank you for your attention!