The Bureau of Meteorology

MultiSiteBoost: Applying XGBoost to Site-specific Weather Forecast

Mengmeng HAN, 08/Jun/2023

Background: Sub-grid-scale Variation

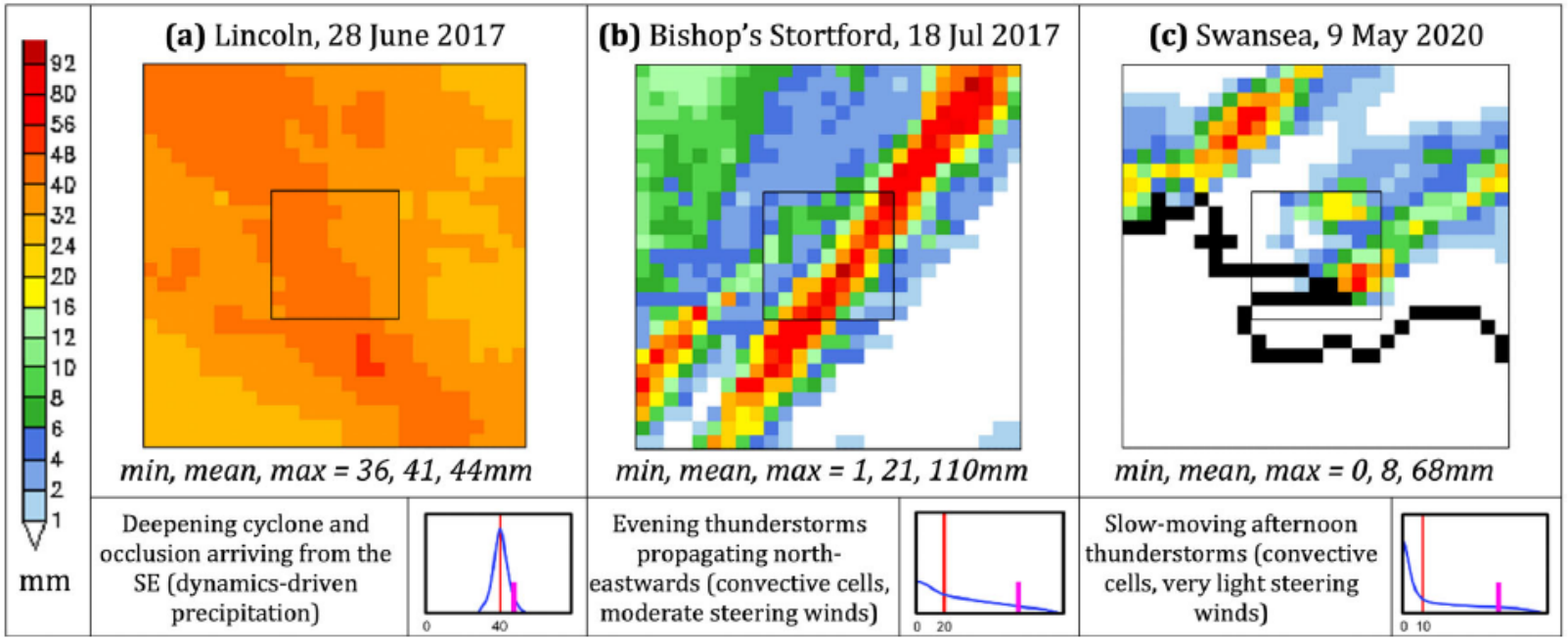
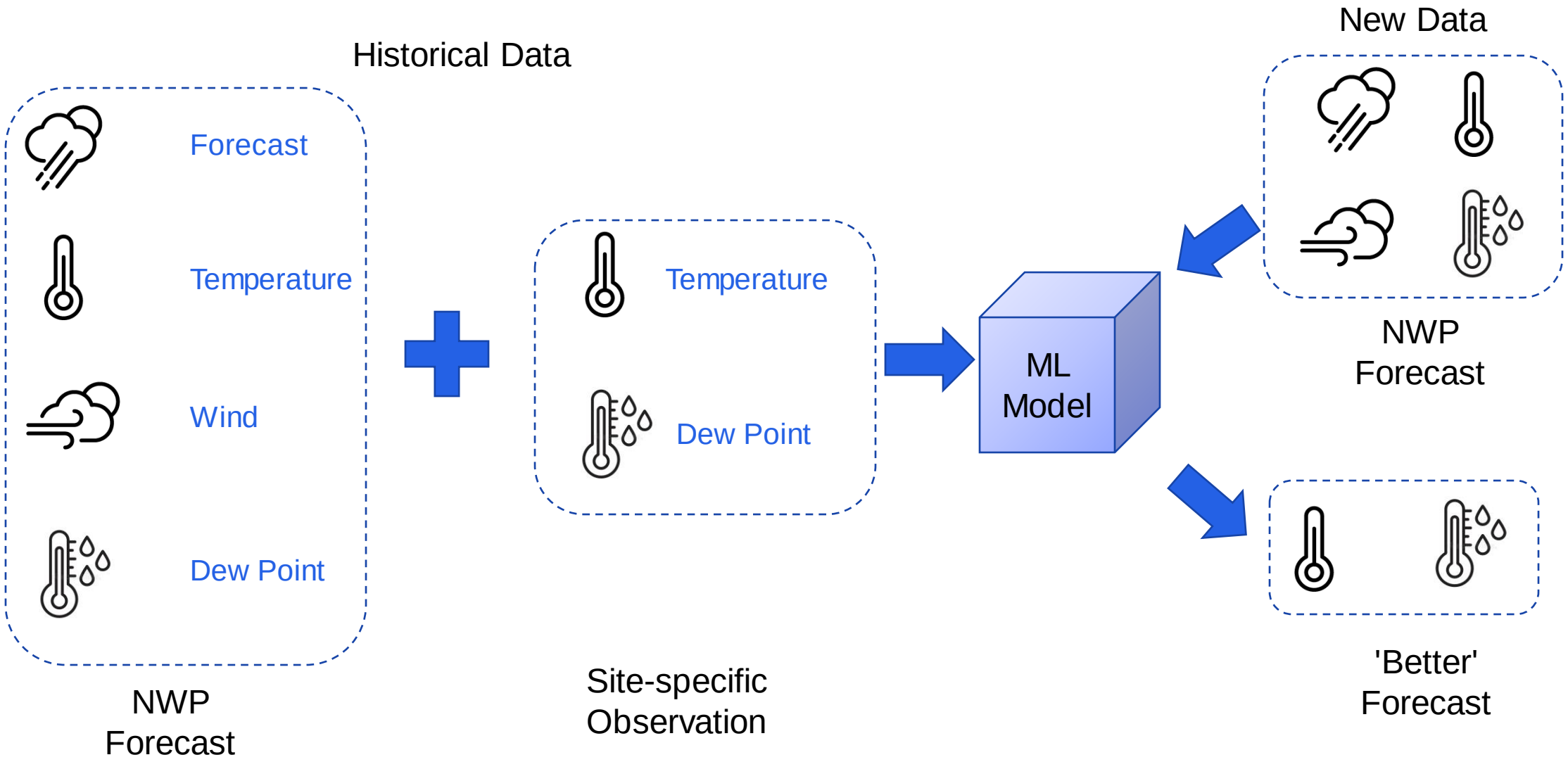


Fig. 1 Three cases of 24 h radar-derived rainfall totals (mm) in the UK illustrating different types of sub-grid variability. a–c Each denote a different case: cells measure 2 × 2 km, black denotes coasts, full frames are 54 × 54 km; legend for 24 h rainfall (mm) applies to all. Central black boxes denote an ECMWF ensemble gridbox (18 × 18 km), for which minimum, mean, and maximum rainfall is shown beneath. Named locations lie approximately mid-panel; all are in regions where relatively flat topography makes radar-derived totals more reliable. Bottom row explains the synoptic situations; inset graphs show, conceptually, how a raw ensemble member forecast (red) for the box should be converted by ecPoint into a probability density function (PDF) for point values (blue) within the box; pink line denotes the 95th percentile; *x*-scale is linear. Flash floods affected the two regions with red pixel clusters in (**c**)[64]. Radar images are from *netweather.tv*.

Hewson, Timothy David, and Fatima Maria Pillosu. "A low-cost post-processing technique improves weather forecasts around the world." *Communications Earth & Environment* 2.1 (2021): 132.

Scope of Work

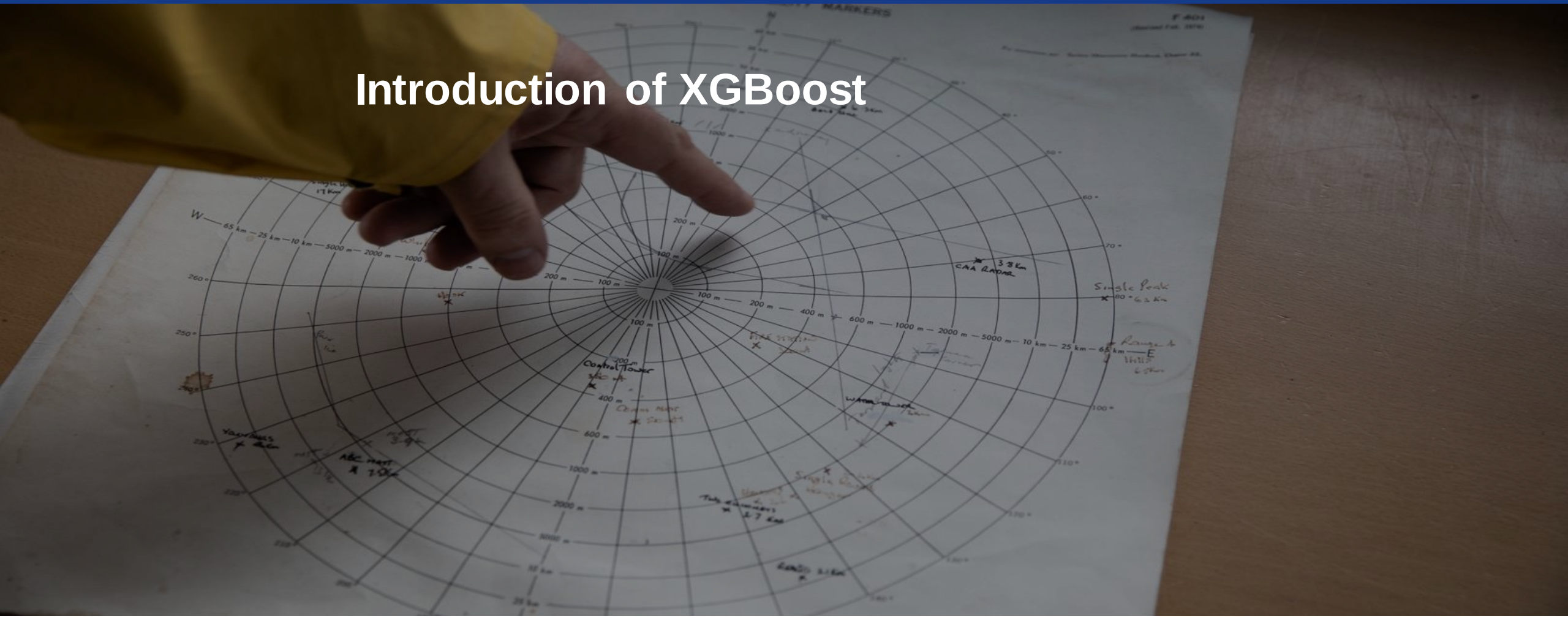


The Bureau of Meteorology

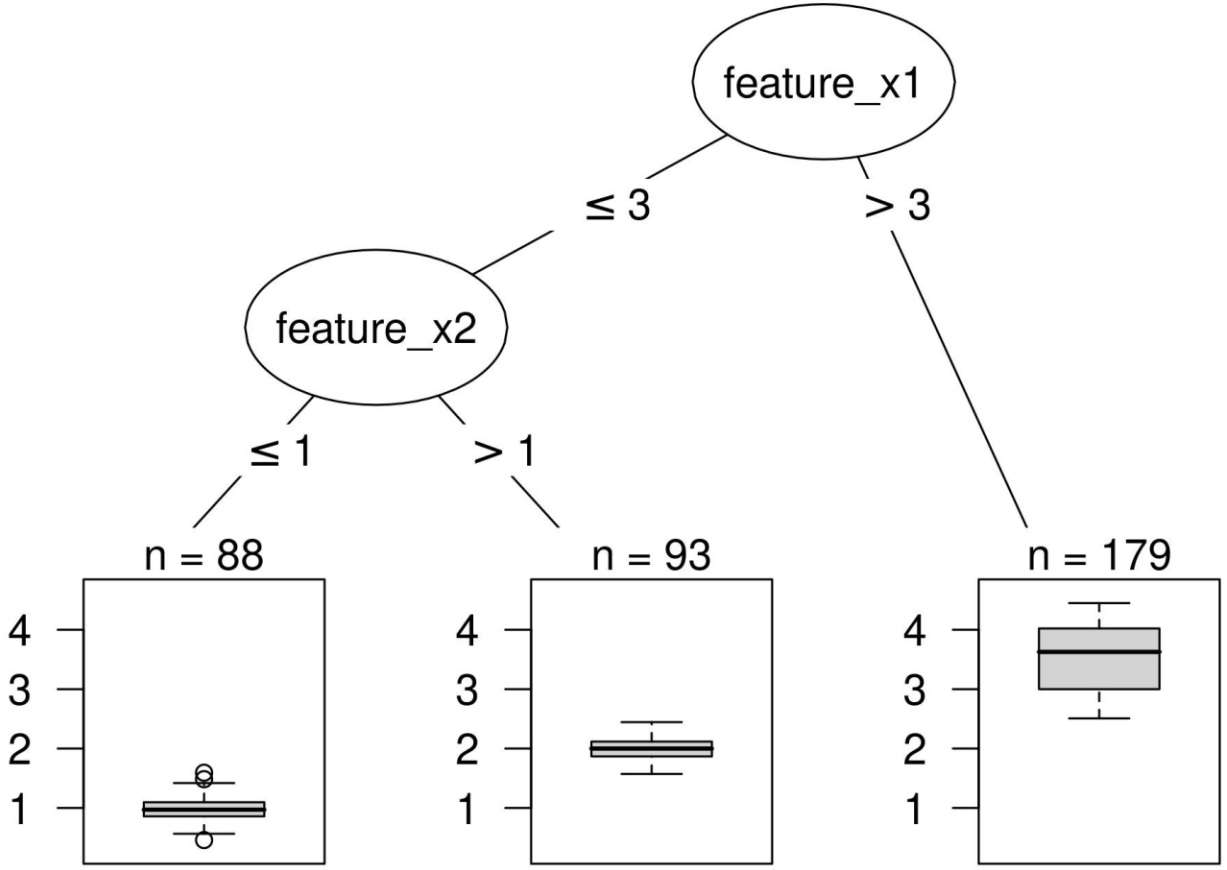
Presentation Outline

- Introduction of XGBoost
- MultiSiteBoost: Data Source
- MultiSiteBoost: Data Preparation and Modelling
- MultiSiteBoost: Verification
- MultiSiteBoost: Results
- Future development

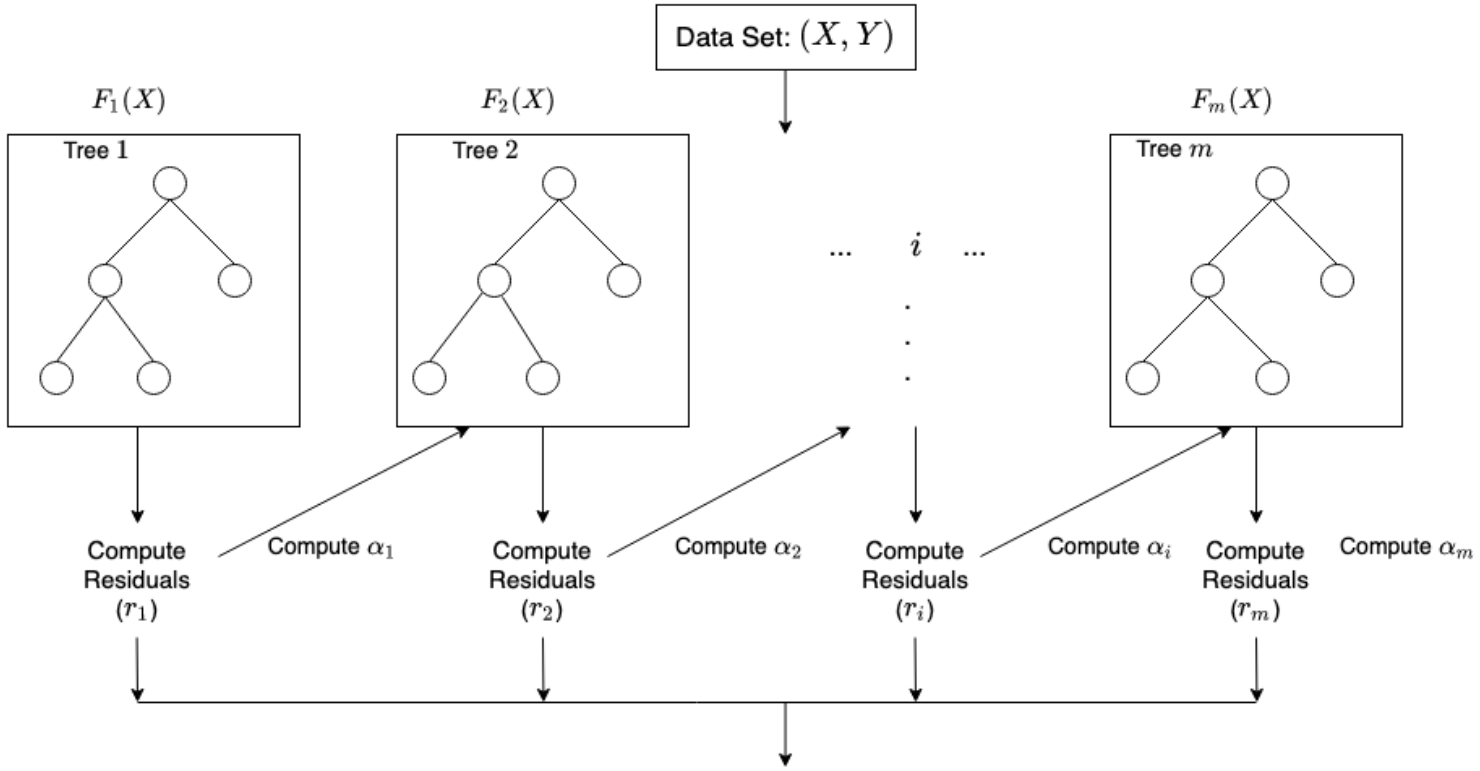
Introduction of XGBoost



Decision Tree: Basics



Decision Tree: XGBoost

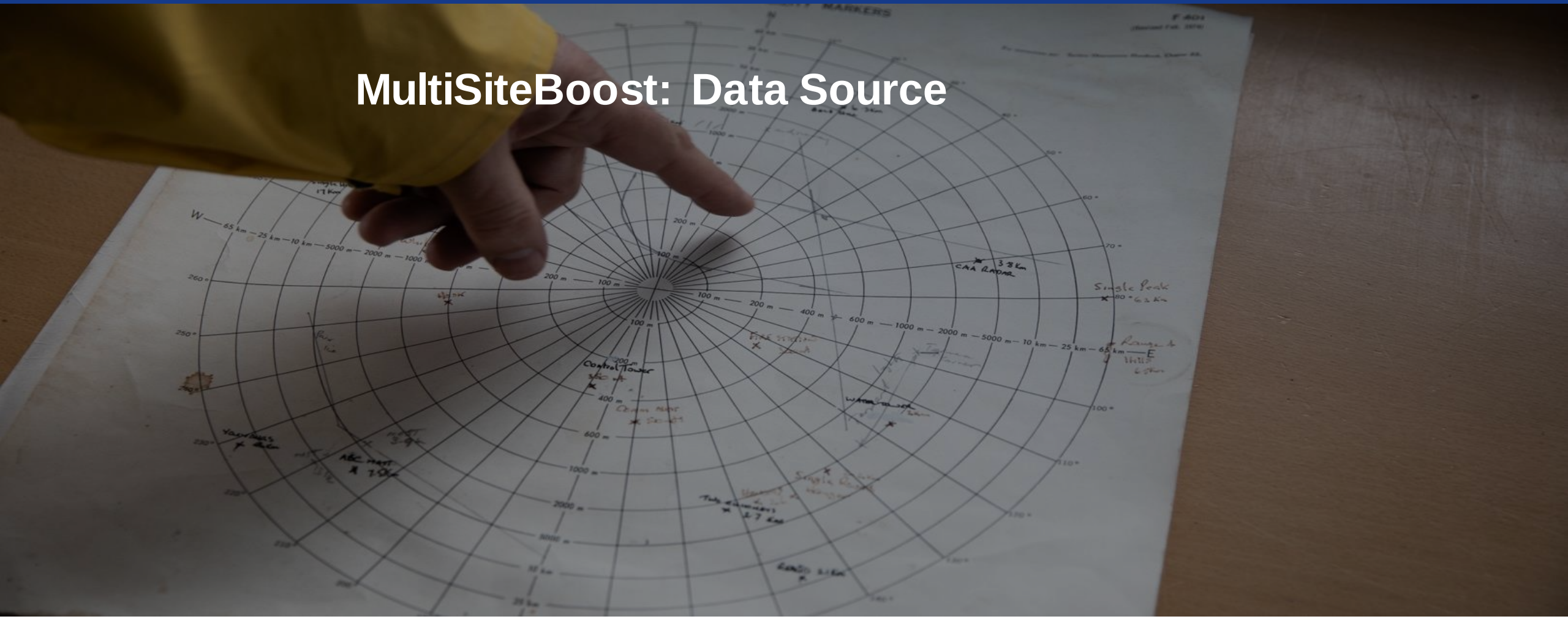


$$F_m(X) = F_{m-1}(X) + \alpha_m h_m(X, r_{m-1}),$$

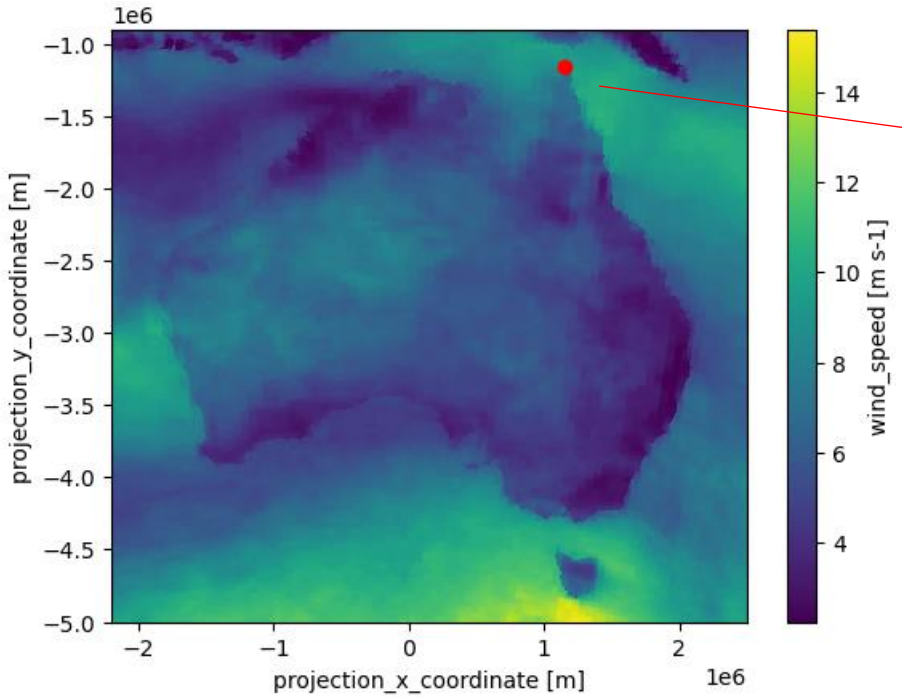
where α_i, and r_i are the regularization parameters and residuals computed with the i^{th} tree respectfully, and h_i is a function that is trained to predict residuals, r_i using X for the i^{th} tree. To compute α_i we use the residuals computed, r_i and compute the following: $arg \min_{\alpha} = \sum_{i=1}^{m} L(Y_i, F_{i-1}(X_i) + \alpha h_i(X_i, r_{i-1}))$ where $L(Y, F(X))$ is a differentiable loss function.

- **Further Reading**:
  - Why do tree-based models still outperform deep learning on typical tabular data?, NeurIPS 2022 Track Datasets and Benchmark

MultiSiteBoost: Data Source



Data Extraction



NW		N		NE
		….		
W		SITE Center		E
		….		
SW		S		SE

x y

Valid_time	temperature_at_screen_level	temperature_of_dew_point_at_screen_level	…	TMP
31/08/2022 2:00	21.0297	15.2953	…	21.7
31/08/2022 3:00	21.45938	15.45156	…	22.1
31/08/2022 4:00	20.70938	15.49844	…	22.6
31/08/2022 5:00	19.88906	15.51406	…	21.8
31/08/2022 6:00	19.07656	15.5297	…	20.8

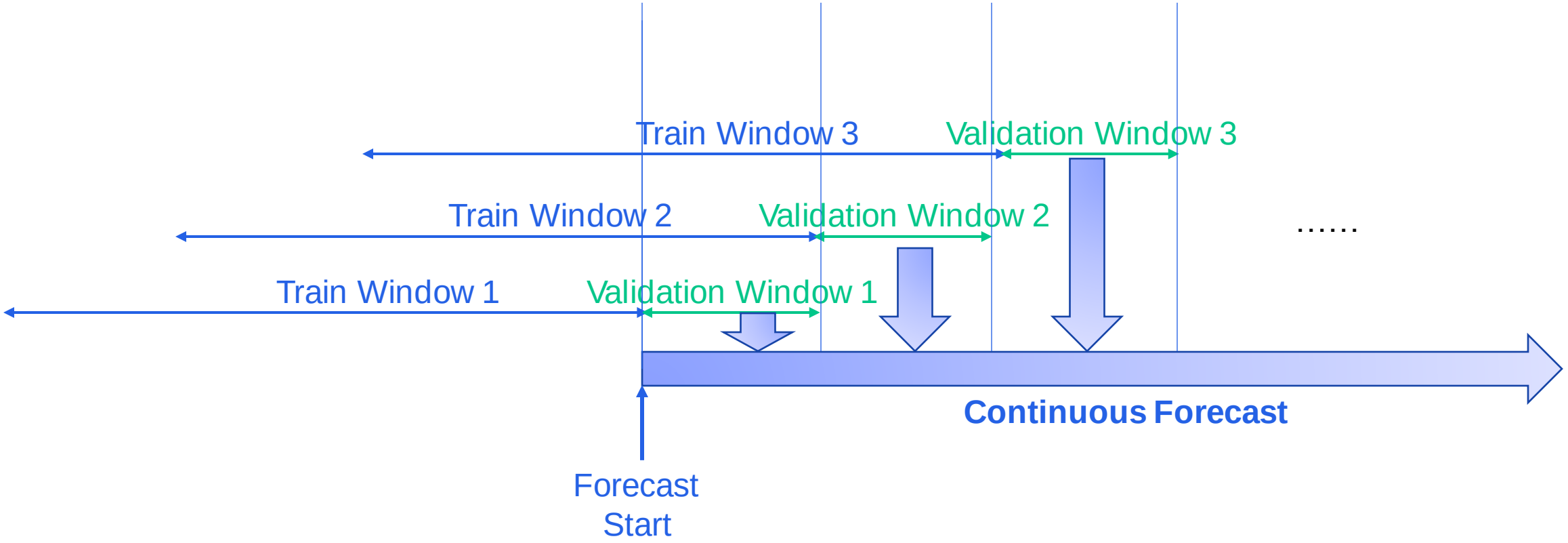
- IMPROVER grid: 4.8 km at screen level
- Hourly data
- Expected values

- Extract: site location and nearby locations 5 grids away

- Tabulate: add observation data at the same validation time
- Observation data source: NOAA ISD, Bureau site obs

MultiSiteBoost: Model

Sliding Window Approach

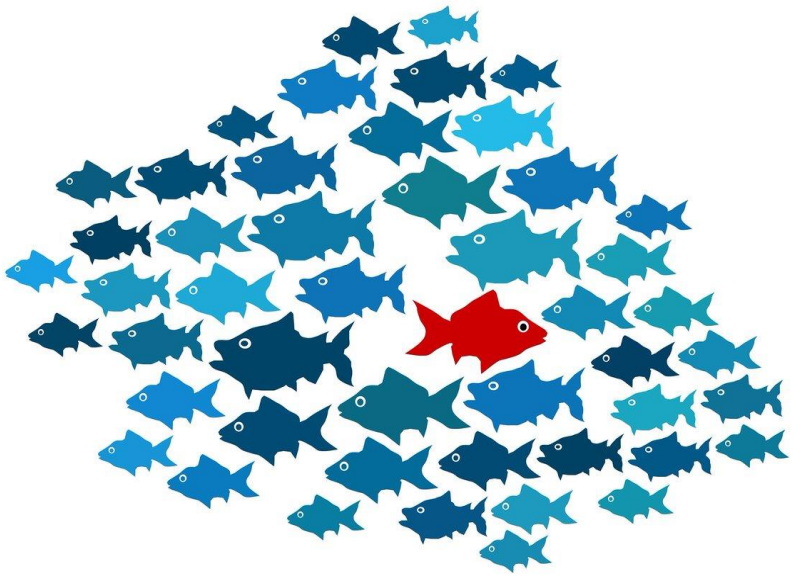


Further References:

- https://machinelearningmastery.com/time-series-forecasting-supervised-learning
- Chandar, B. S., Rajagopalan, P., & Ranganathan, P. (2023, March). Short-Term AQI Forecasts using Machine/Deep Learning Models for San Francisco, CA. In 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC) (pp. 0402-0411). IEEE.

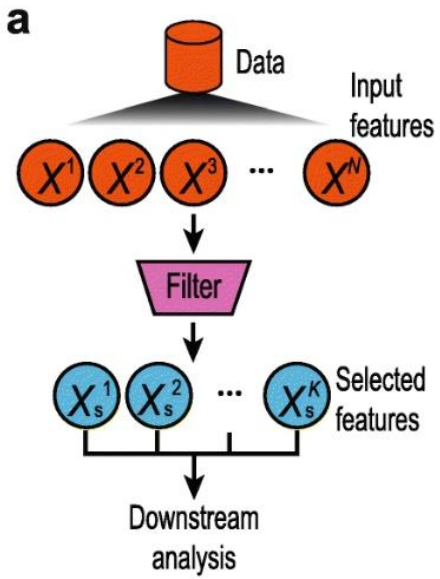
Optimization

Handling Outlier Data



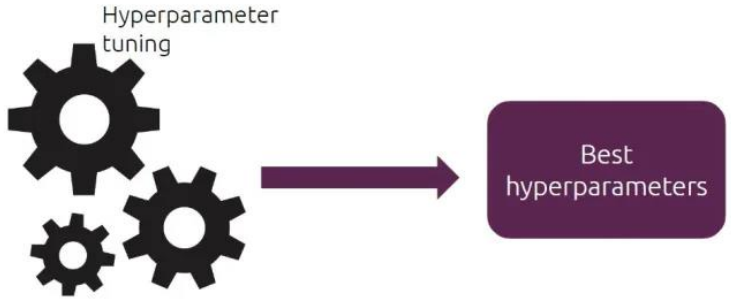
[Detecting and Treating Outliers | How to Handle Outliers (analyticsvidhya.com)]

Feature Selection



Yang, P., Huang, H. & Liu, C. Feature selection revisited in the single-cell era. *Genome Biol* **22**, 321 (2021). https://doi.org/10.1186/s13059-021-02544-3

Hyperparameter Tuning



Examples of hyperparameters in XGBoost:

- max_depth
- learning_rate
- n_estimators
- colsample_bytree

Hyperparameter tuning for ML models | Ubuntu

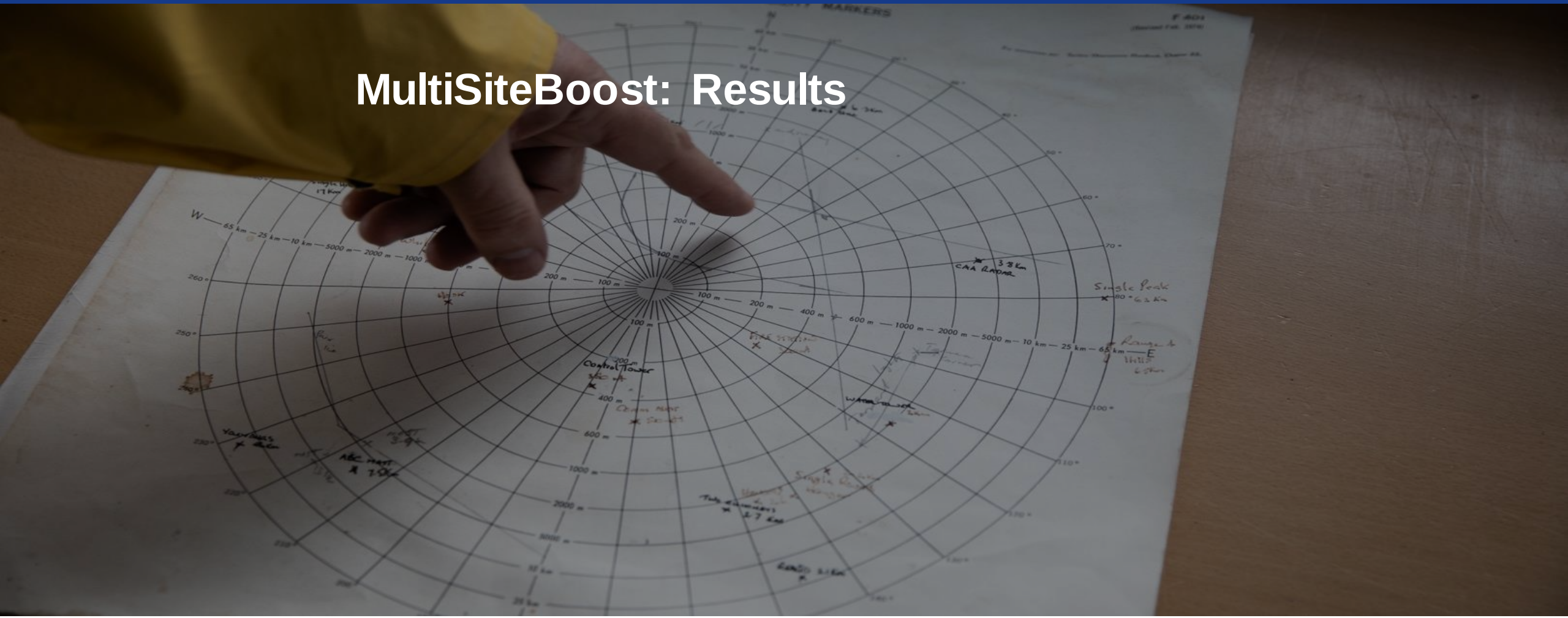
MultiSiteBoost: Verification

Evaluated Metrics

Calculated within Forecast Period, for each lead day

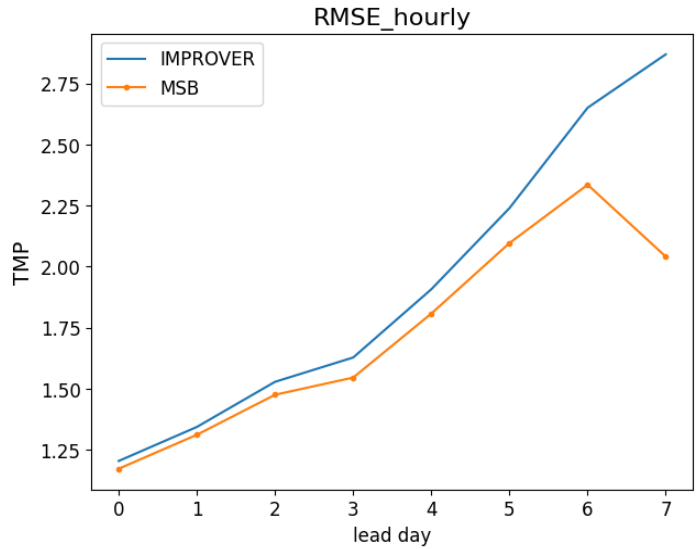
- **Hourly RMSE/MAE (Primary)**
- Daily max RMSE/MAE
- Daily min RMSE/MAE
- Percentage of days MultiSiteBoost generates more accurate daily max/min
- Percentage of forecast data points with error > 2.0 deg
- 90-th percentile of error

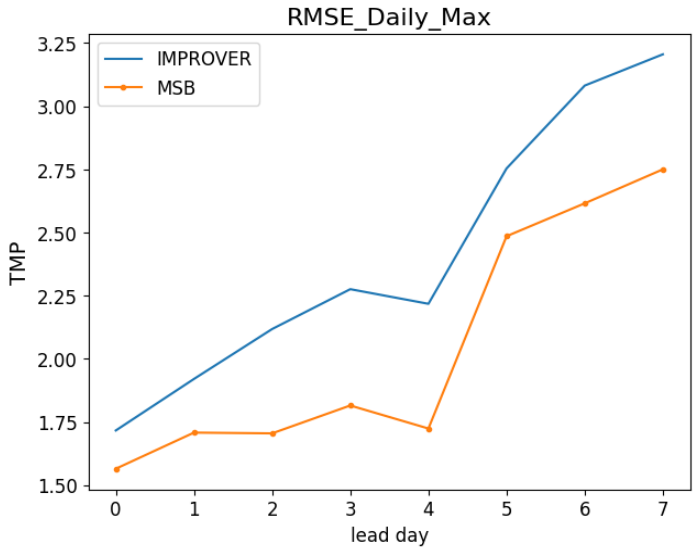
MultiSiteBoost: Results

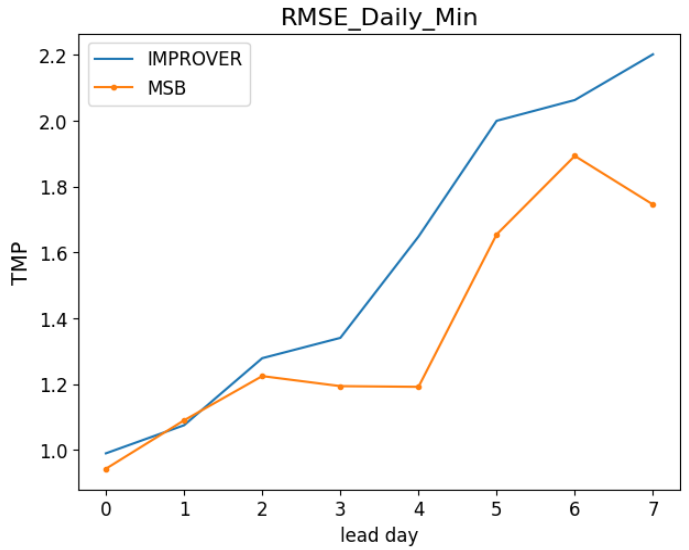


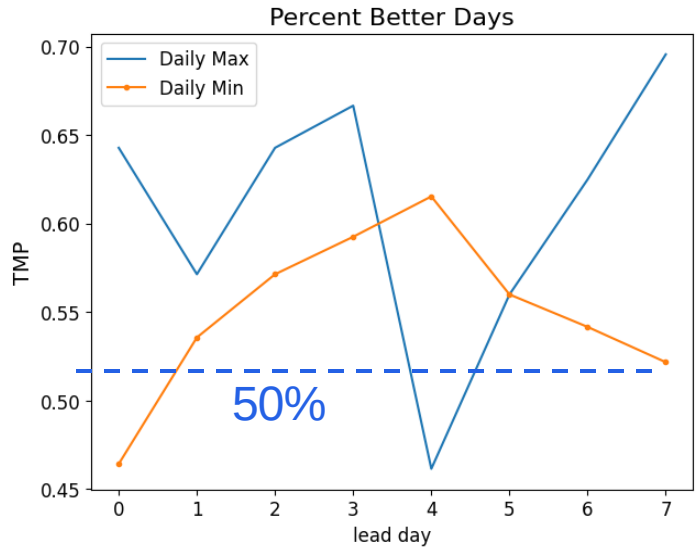
Metrics in Line Plots

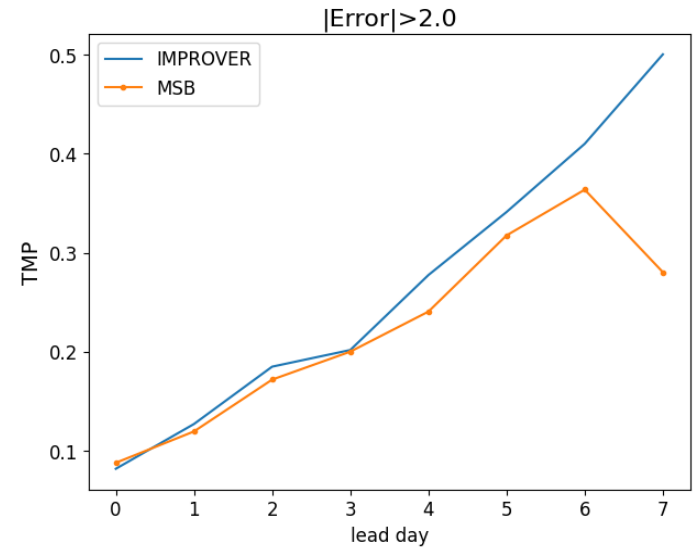
Site: HOBART ELLERSLIE ROAD, Temperature, 2022-10-27 – 2023-11-23

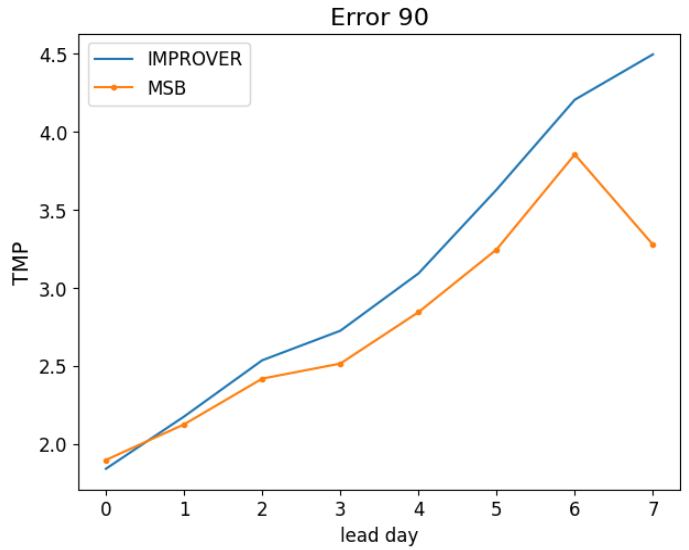






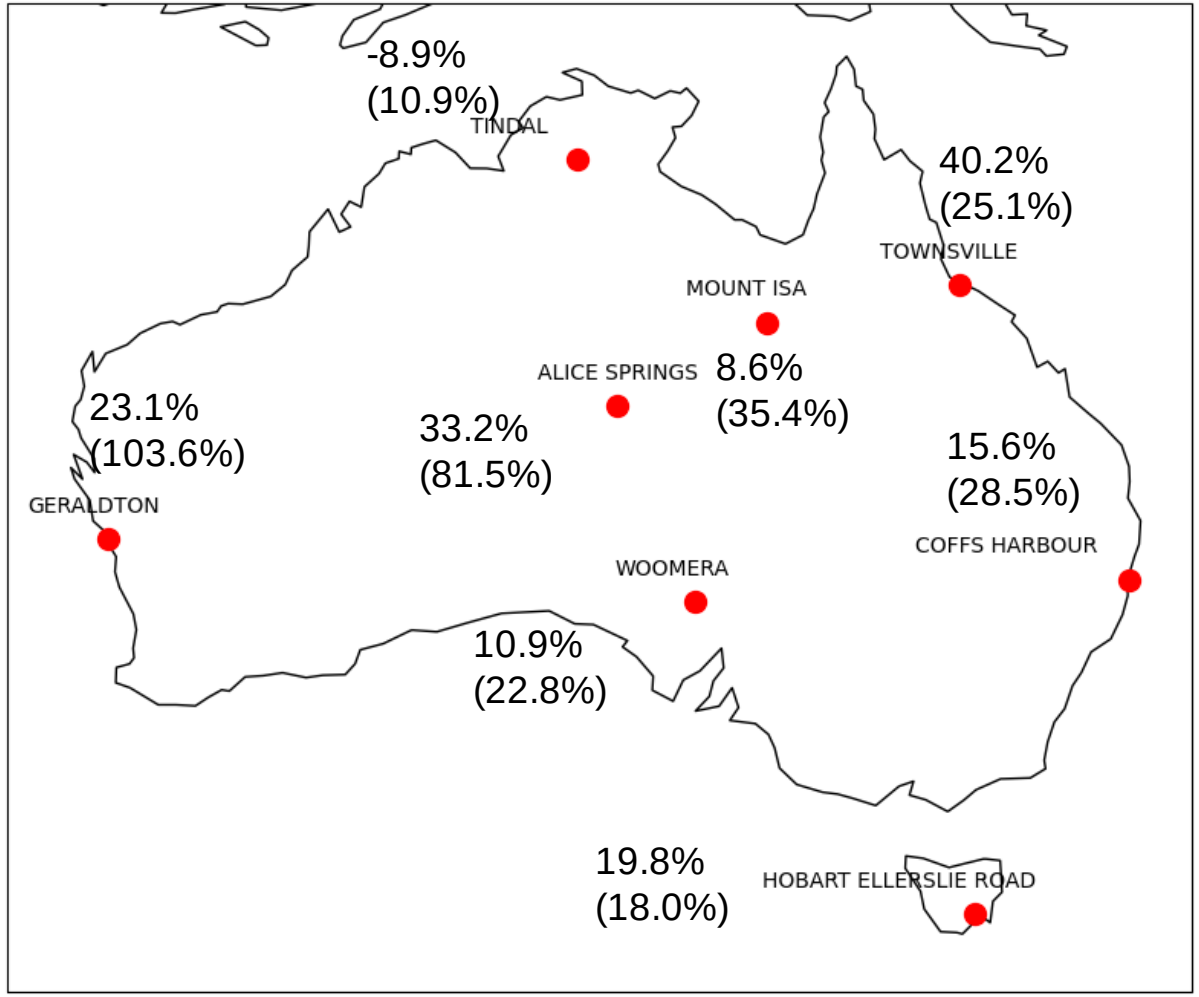




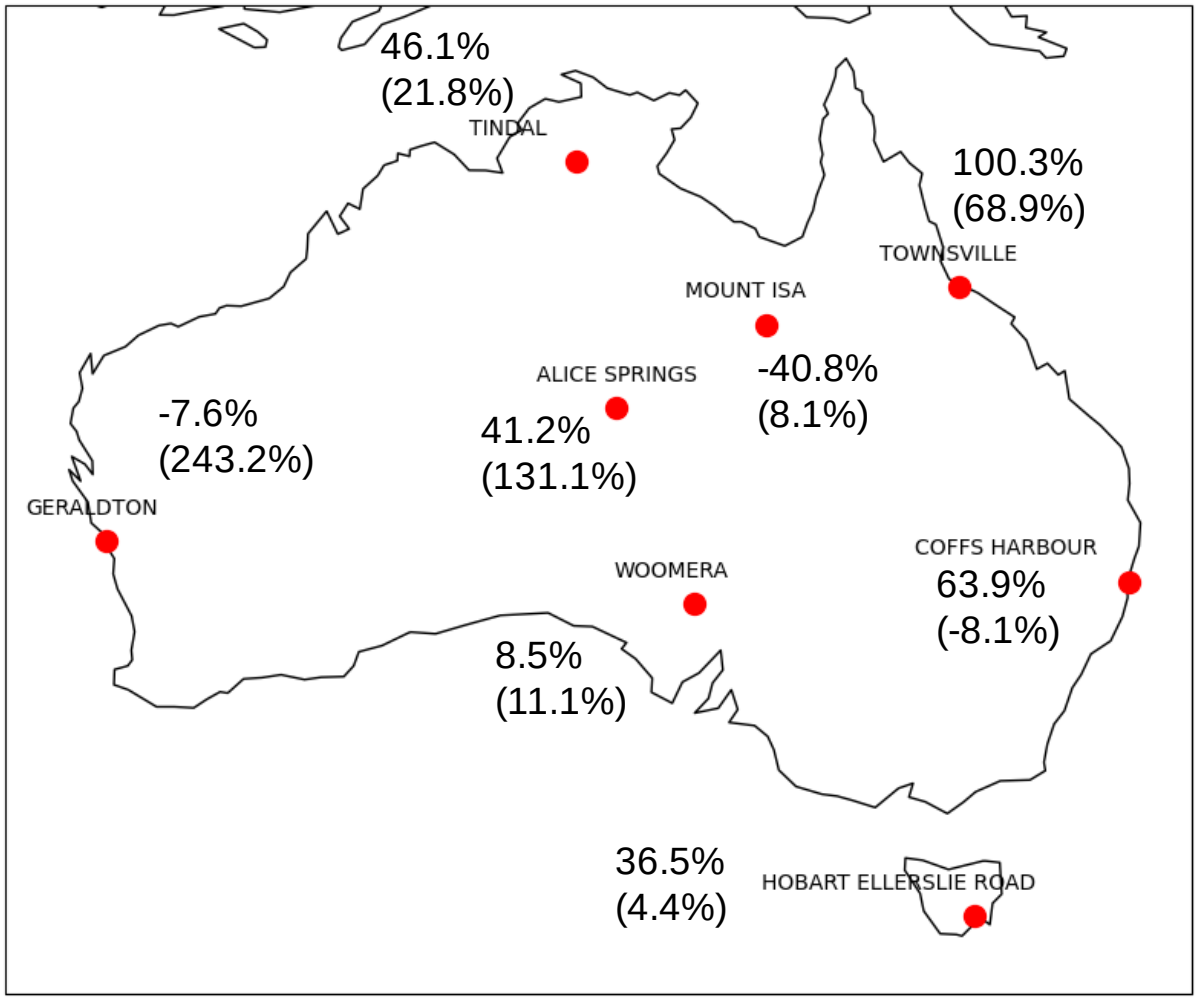


Overall Metrics

RMSE Hourly, TMP (DEW)



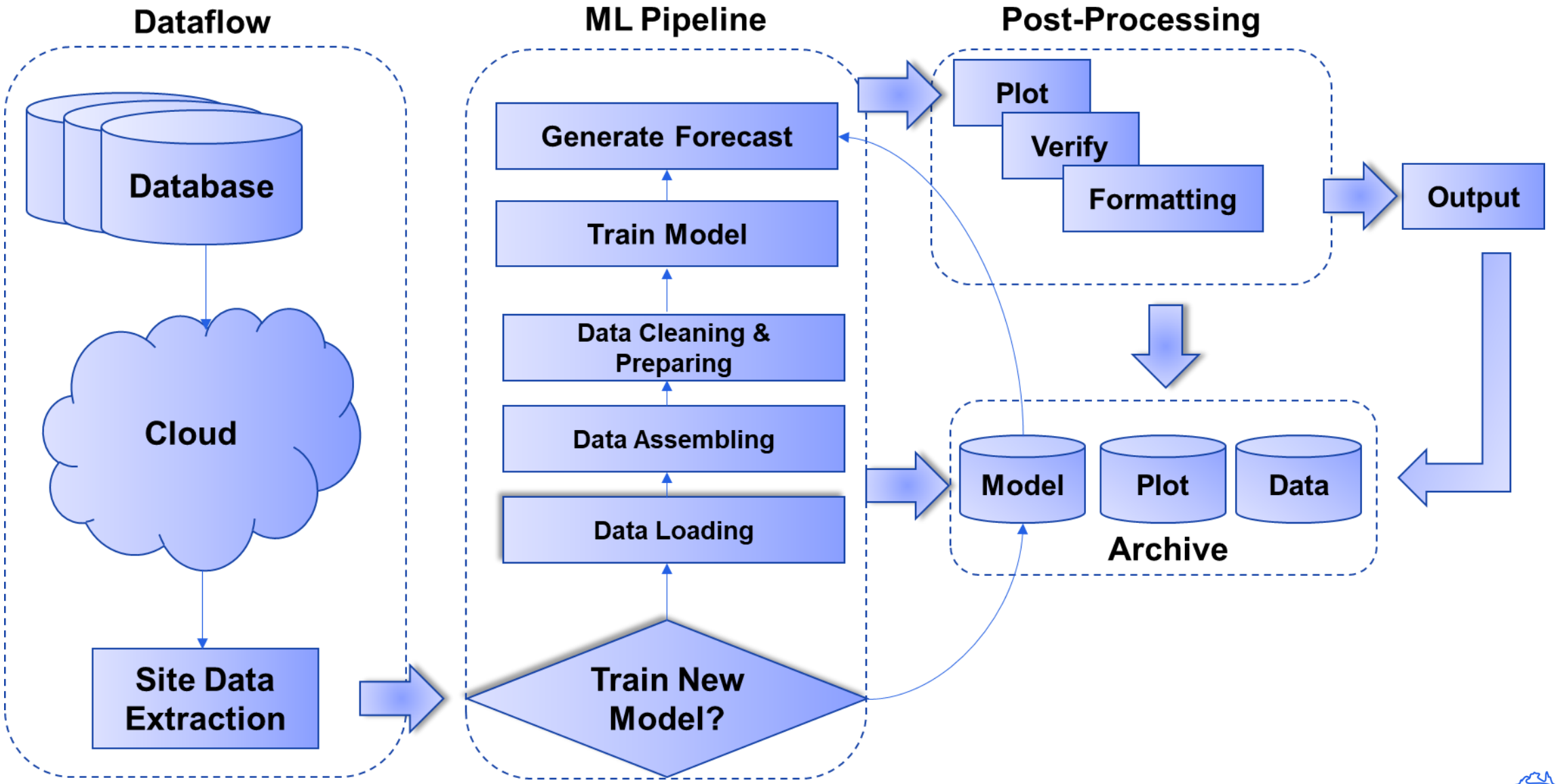
RMSE Daily Max TMP (DEW)



- Average gain of a metric compared with IMPROVER across all lead days, positive means better

Future Development

MultiSiteBoost Architecture Diagram



The Bureau of Meteorology

Thank you

Mengmeng HAN

Add contact details or website here if needed