Met Office

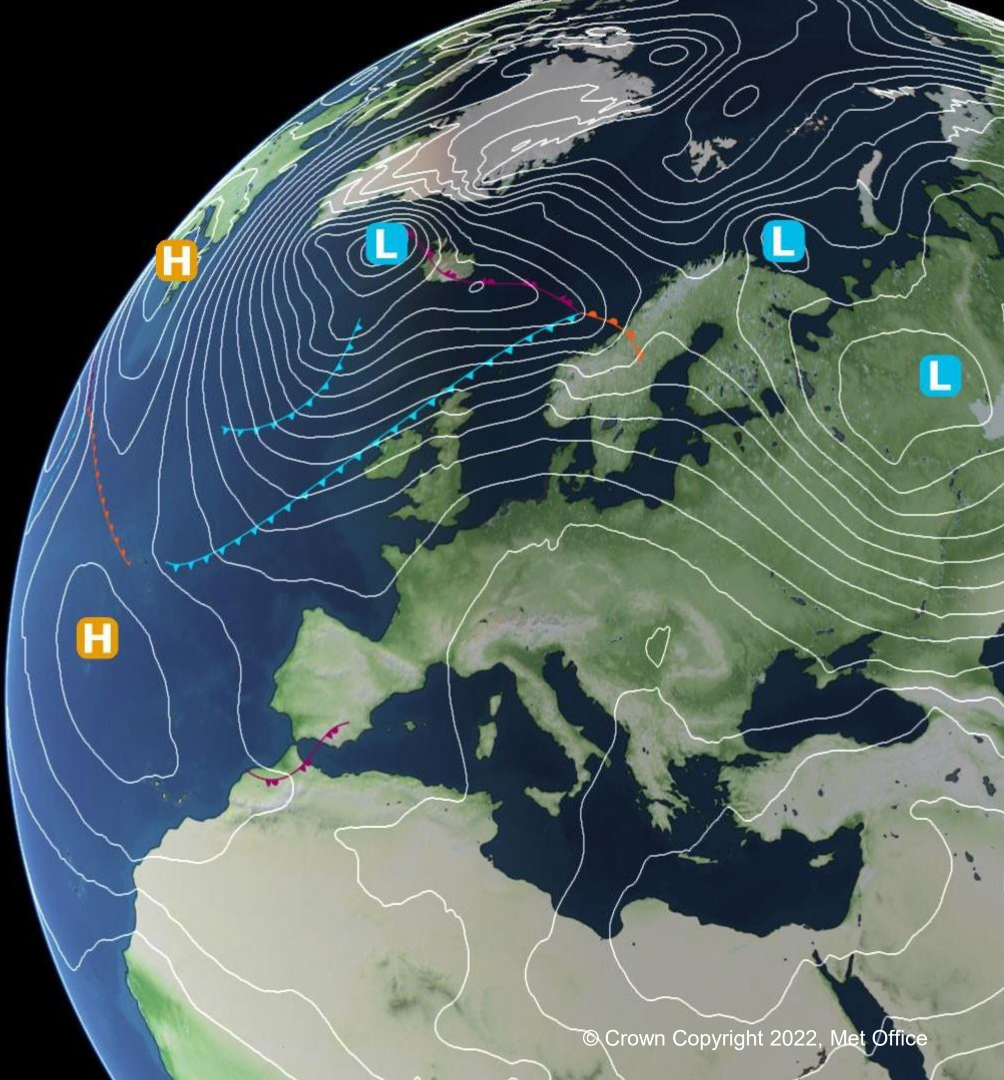
Analysis of MOGREPS-UK ensemble spread and skill and their prediction using machine learning

Carlo Cafaro

Research to Operations, Met Office

Norrköping, Sweden: 22 – 25 September 2025

47th EWGLAM and 32nd SRNWP workshop



Met Office

Background motivation

- MOGREPS-UK has been running in its actual configuration since 2019. Lack of spread has been a proved issue, confirmed both by data and operational meteorologists (e.g. **Porson et al, 2020**)

- First motivation of this short study was to investigate the evolution of spread/skill relationship since 2019 until today (or until the data are available): has it improved?

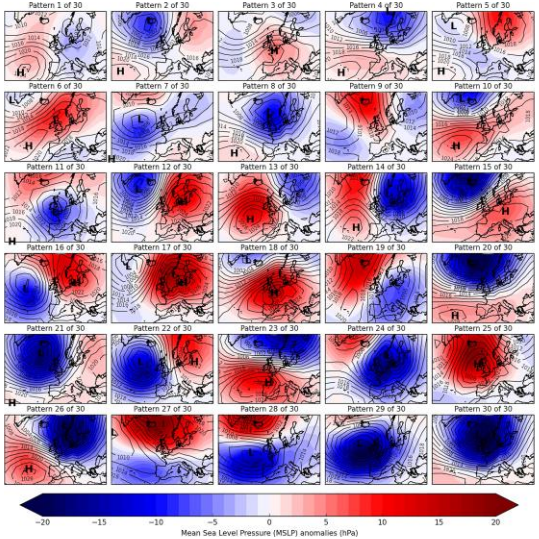
- More recently there has been an increased effort/interest for using ML at Met Office. So I used this spread/skill dataset to train/test some ML methods as proof of concept of predicting forecast uncertainty using machine learning (**Scher and Messori, 2018**).

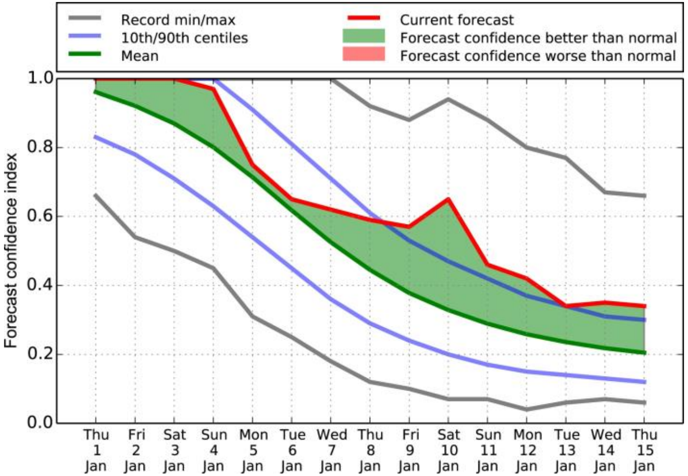
Porson AN, Carr JM, Hagelin S, et al. Recent upgrades to the Met Office convective-scale ensemble: An hourly time-lagged 5-day ensemble. *Q J R Meteorol Soc*. 2020; 146: 3245–3265. **https://doi.org/10.1002/qj.3844**

Scher S, Messori G. Predicting weather forecast uncertainty with machine learning. Q J R Meteorol Soc. 2018; 144: 2830–2841. **https://doi.org/10.1002/qj.3410**

Research question

Can we predict ensemble spread and/or skill using large-scale predictors with Machine Learning ?





Decider weather regimes

Confidence Forecast index

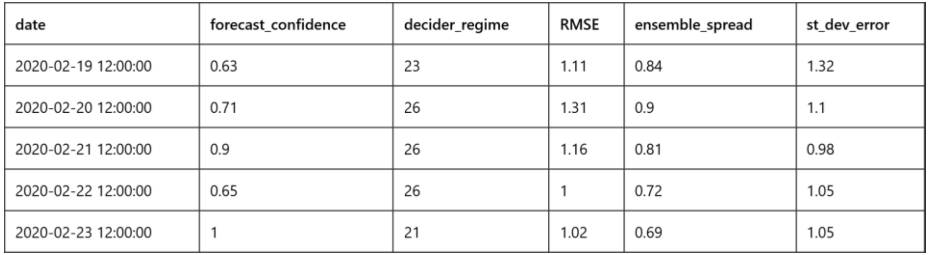
(Neal et al., 2016)

Neal, R., Fereday, D., Crocker, R. and Comer, R.E. (2016), A flexible approach to defining weather patterns and their application in weather forecasting over Europe. Met. Apps, 23: 389-400. https://doi.org/10.1002/met.1563

Met Office

Preparing the Dataset

- **ensemble spread and skill** (RMSE, st dev error [of control fcst])
  (Apr 2019 – Dec 2023)
  hourly from MOGREPS-UK (up to T+120h)

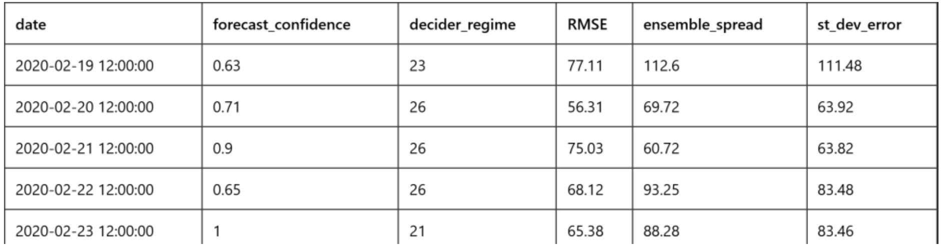


date	forecast_confidence	decider_regime	RMSE	ensemble_spread	st_dev_error
2020-02-19 12:00:00	0.63	23	1.11	0.84	1.32
2020-02-20 12:00:00	0.71	26	1.31	0.9	1.1
2020-02-21 12:00:00	0.9	26	1.16	0.81	0.98
2020-02-22 12:00:00	0.65	26	1	0.72	1.05
2020-02-23 12:00:00	1	21	1.02	0.69	1.05

- **Forecast confidence index**
  (Feb 2020 – today)
  daily from MOGREPS-G (up to 7 days)

→ Daily data
(Feb 2020 – Dec 2023)

- **decider regimes**
  (Jan 2009 – today)
  daily From Global Model (up to 7days)



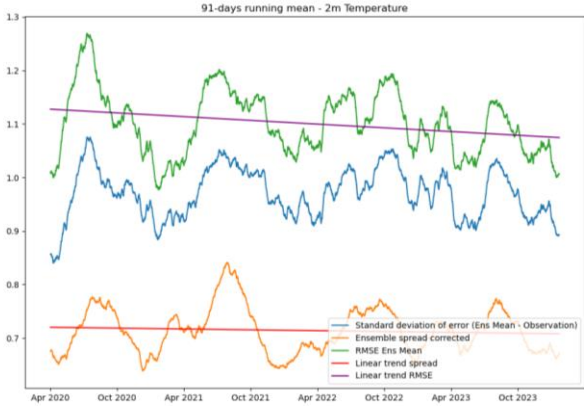
date	forecast_confidence	decider_regime	RMSE	ensemble_spread	st_dev_error
2020-02-19 12:00:00	0.63	23	77.11	112.6	111.48
2020-02-20 12:00:00	0.71	26	56.31	69.72	63.92
2020-02-21 12:00:00	0.9	26	75.03	60.72	63.82
2020-02-22 12:00:00	0.65	26	68.12	93.25	83.48
2020-02-23 12:00:00	1	21	65.38	88.28	83.46

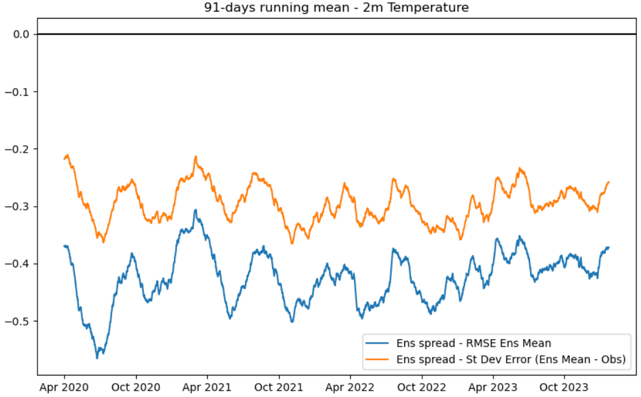
Data courtesy of Rachel North (spread/skill) and Robert Neal (decider, forecast confidence)

2m Temperature – spread/skill timeseries

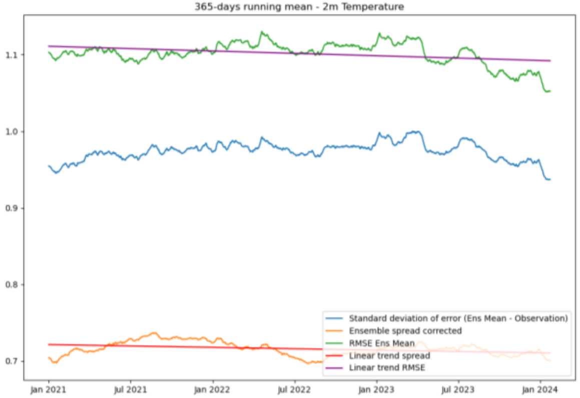
Differences

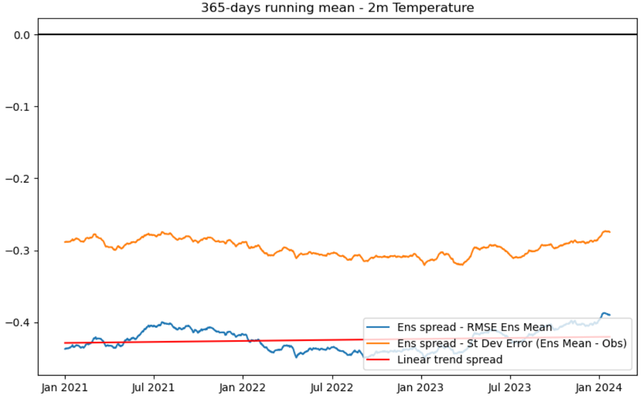
3 months running mean





12 months running mean



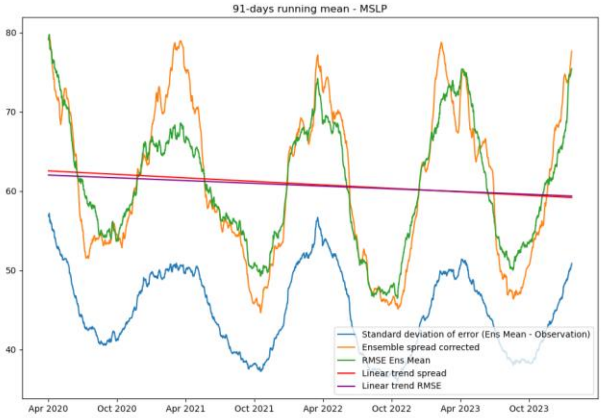


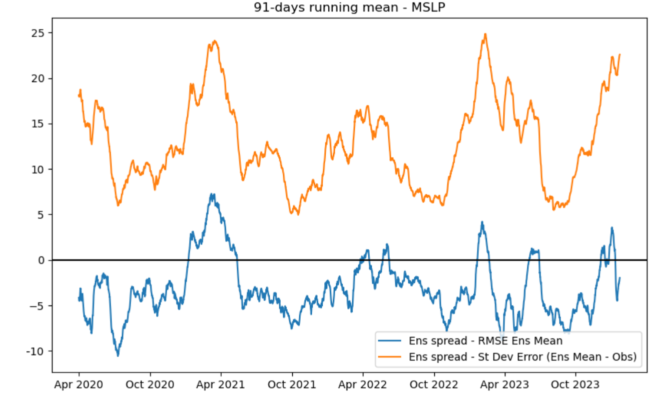
- Spread and RMSE both declining, with differences increasing over time

MSLP– spread/skill timeseries

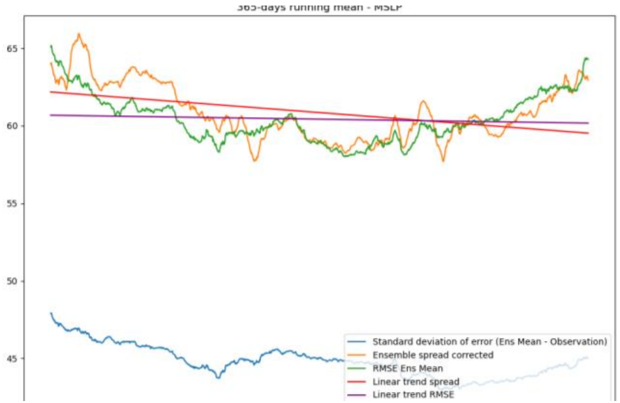
Differences

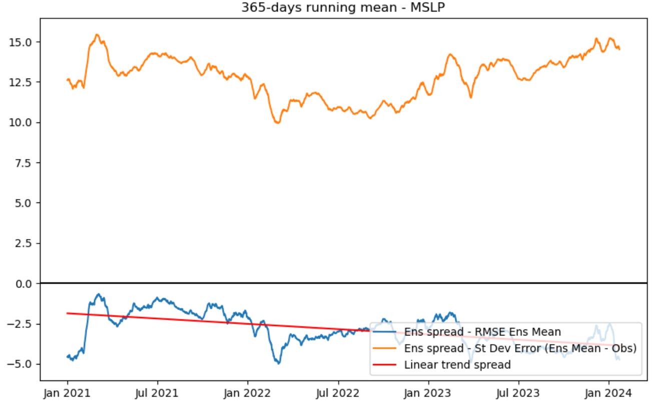
3 months running mean





12 months running mean





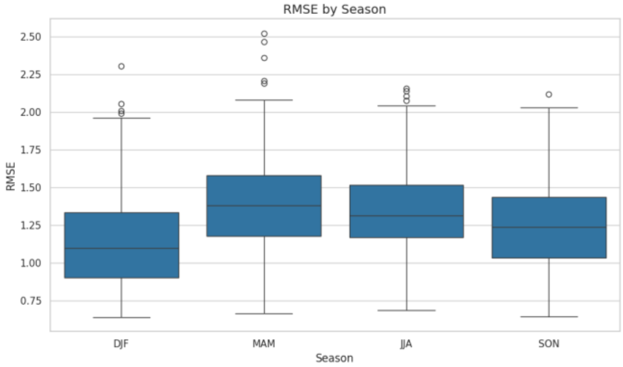
- Spread and RMSE both declining, with differences decreasing over time

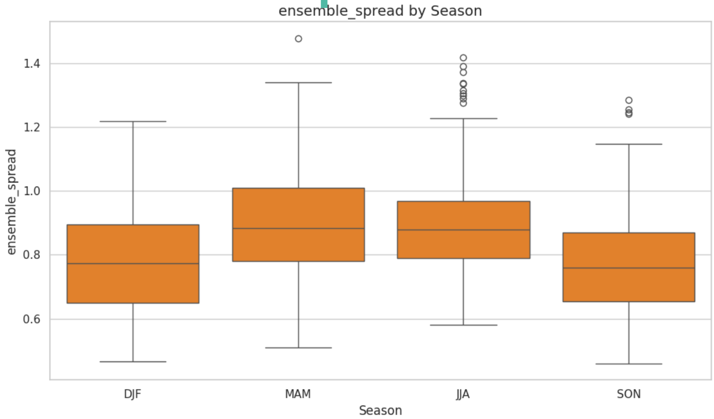
Seasonal dependency

RMSE

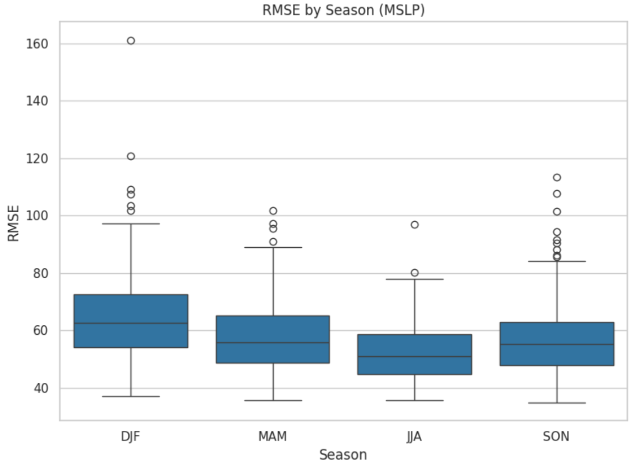
spread

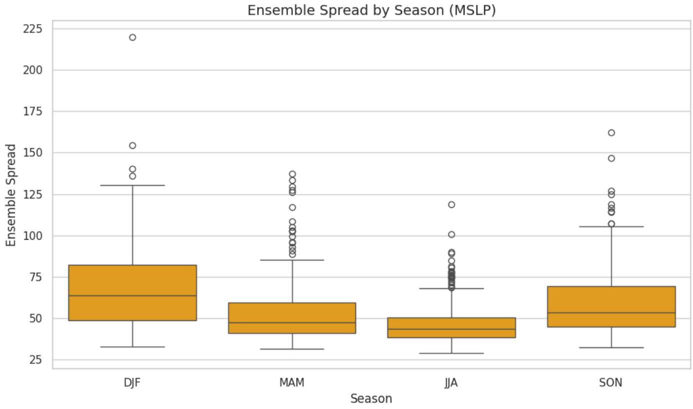
2m Temperature





mslp





First part summary

- Ensemble spread has been calculated including a correction factor to take into account for the finite ensemble size **(Leutbecher and Palmer, 2008).**

- MSLP spread/RMSE peak during winter/autumn season, whereas 2mTemp spread/RMSE have multiple peaks (both summer and spring seasons)

- Negative differences between spread and skill are significant (Wilcoxon test – pvalue<0.05, not shown) implying MOGREPS-UK is significantly underdispersive, at least for these two variables.

- Linear trends are shown to be negative for both the variables and for both spread/RMSE, however the difference trends are different: slightly positive (negative) for 2mT (MSLP).

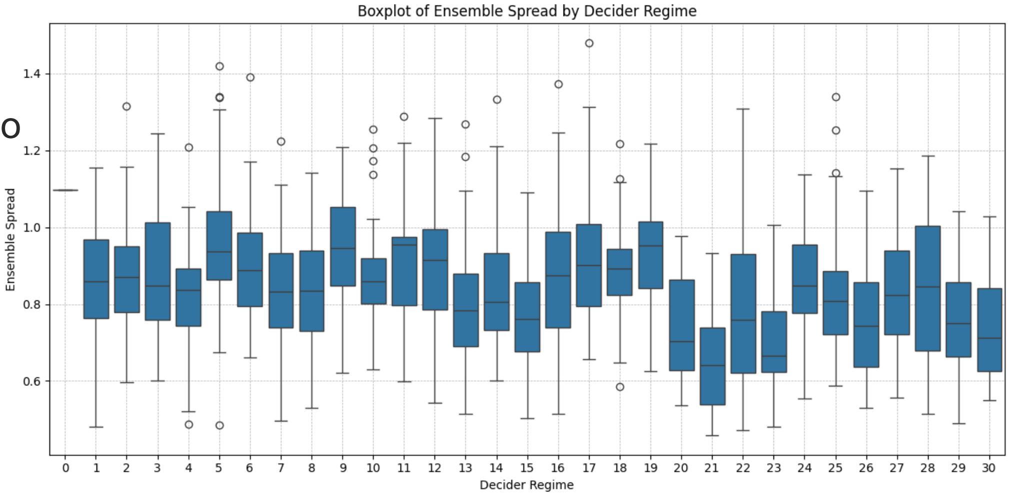
Leutbecher, M. and Palmer, T.N. (2008) Ensemble forecasting. Journal of Computational Physics, 227, 3515–3539

Machine learning prediction

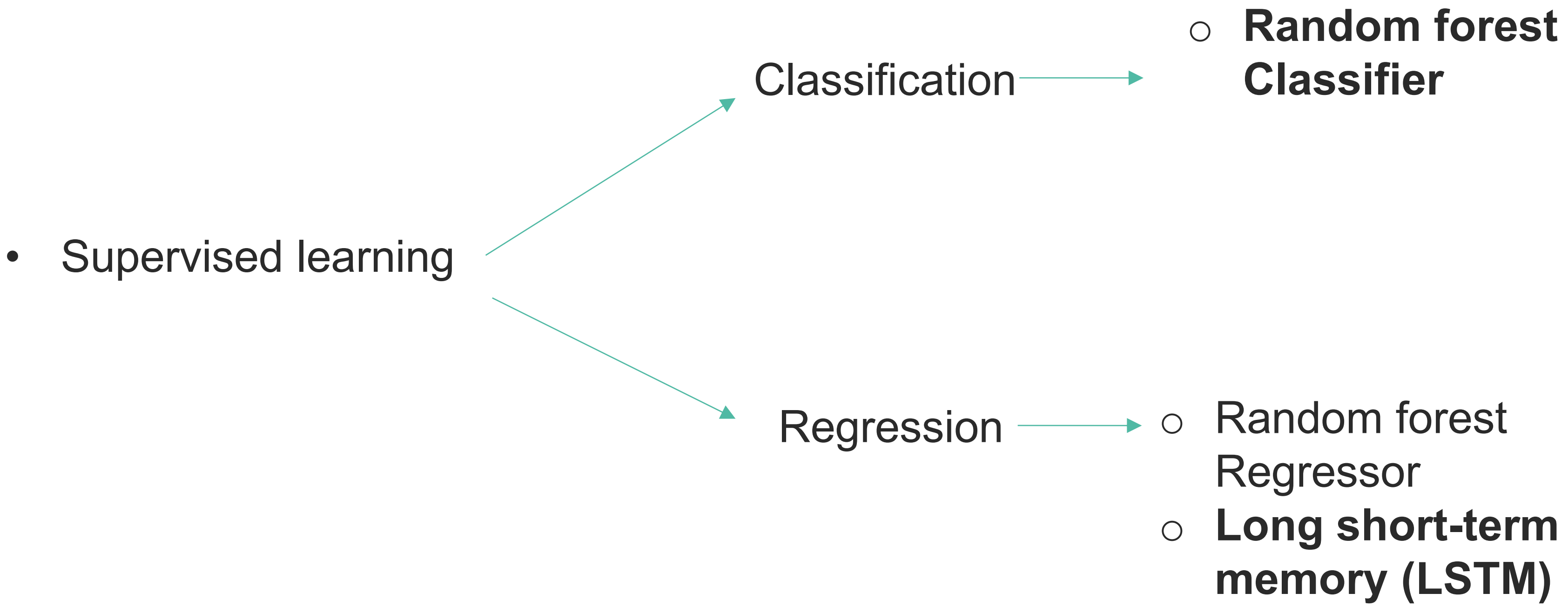
Data wrangling and Experiments design

- Data are 1-D (timeseries) and I had daily data for the features (regimes and confidence forecast index) and hourly data for the ensemble spread/skill. To convert from hourly to daily I took the 24h-median and max of the features.

- **Experiment design**
- Identifying independent and dependent variables
- Which ML method to select?
- Optimising the ML settings



ML methods

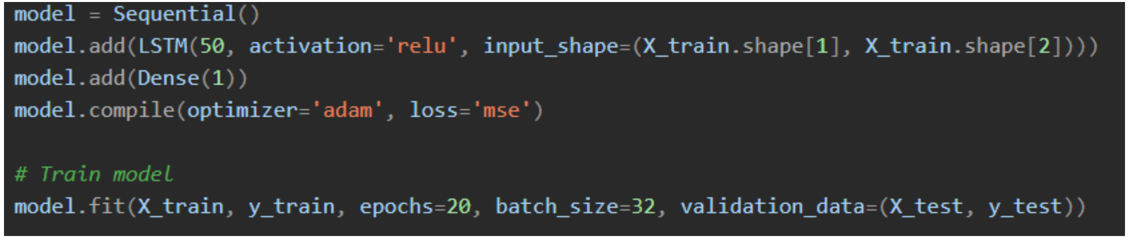


- Supervised learning
  - Classification
    - **Random forest Classifier**
  - Regression
    - Random forest Regressor
    - **Long short-term memory (LSTM)**

Results

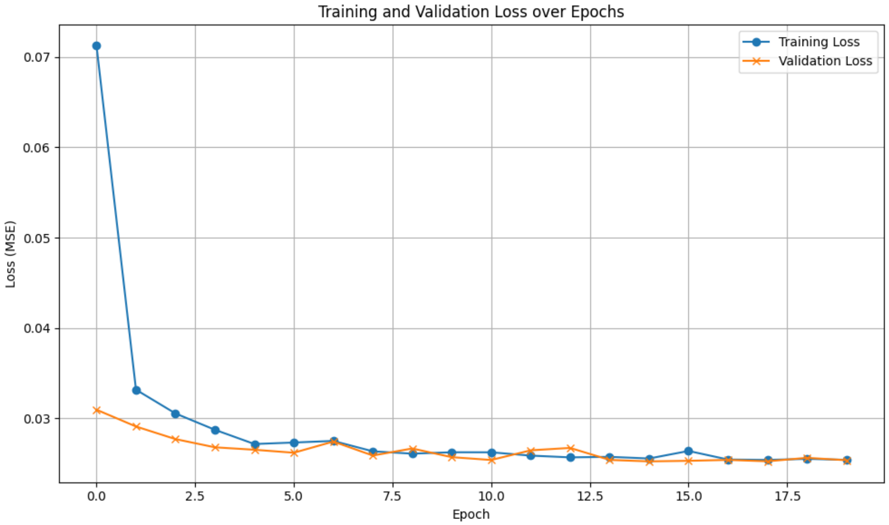
LSTM

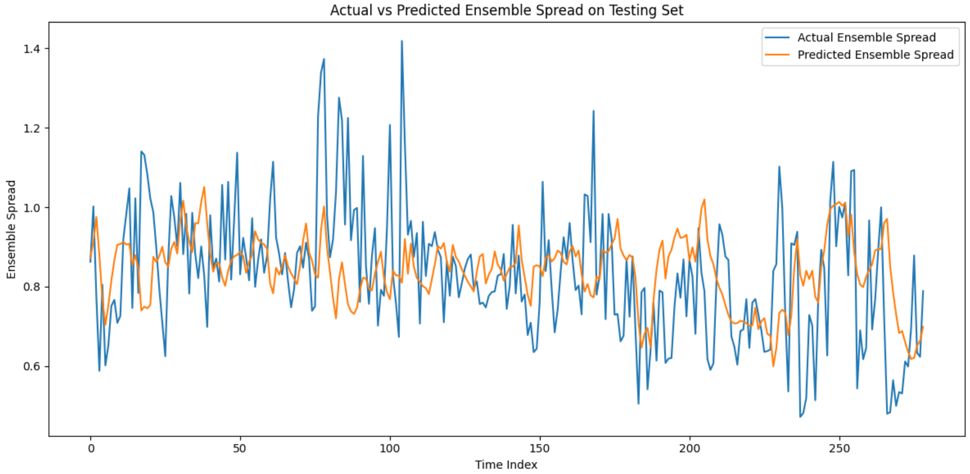
Feature: decider regime/forecast_confidence at times [t-10,t] /
Target: ensemble spread at time t



```
model = Sequential()
model.add(LSTM(50, activation='relu', input_shape=(X_train.shape[1], X_train.shape[2])))
model.add(Dense(1))
model.compile(optimizer='adam', loss='mse')

# Train model
model.fit(X_train, y_train, epochs=20, batch_size=32, validation_data=(X_test, y_test))
```

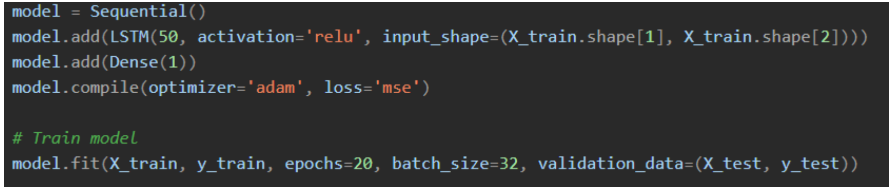




Results

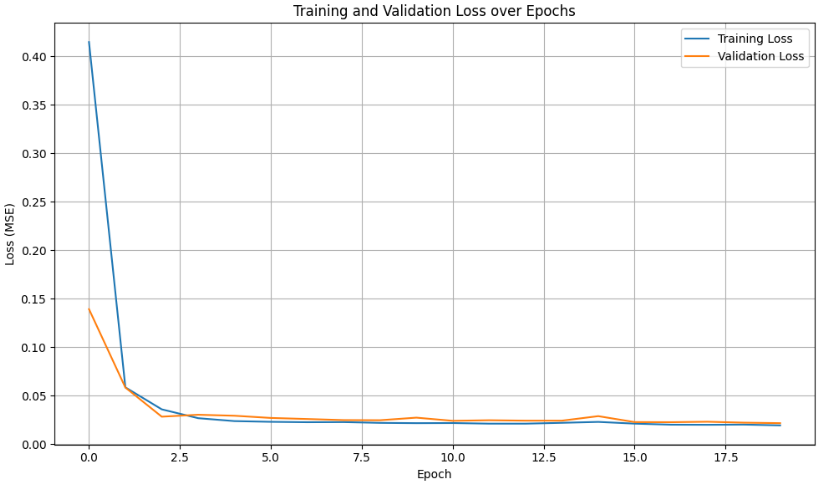
LSTM

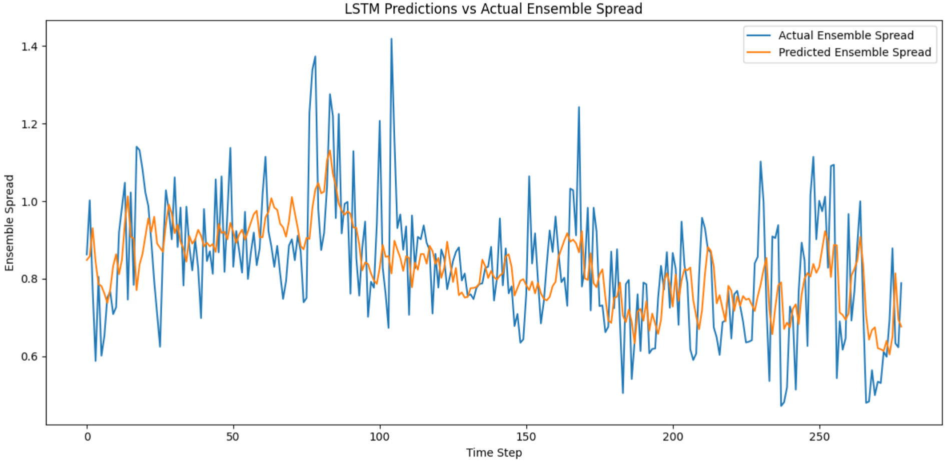
Feature: decider regime/forecast_confidence/day of the year/standard deviation of error at times [t-10,t] / **Target**: ensemble spread at time t



```
model = Sequential()
model.add(LSTM(50, activation='relu', input_shape=(X_train.shape[1], X_train.shape[2])))
model.add(Dense(1))
model.compile(optimizer='adam', loss='mse')

# Train model
model.fit(X_train, y_train, epochs=20, batch_size=32, validation_data=(X_test, y_test))
```

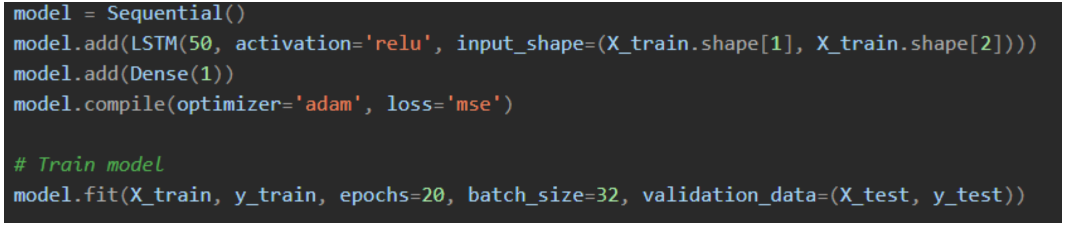




Results

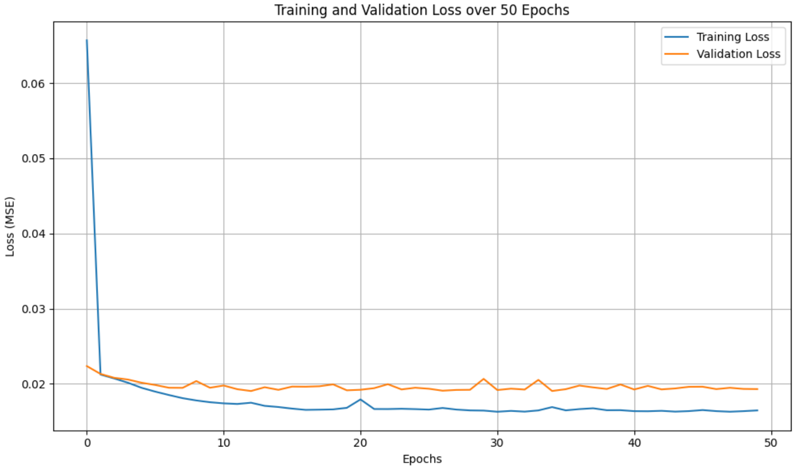
LSTM

Feature: ensemble spread at times [t-10,t-1] /
Target: ensemble spread at time t

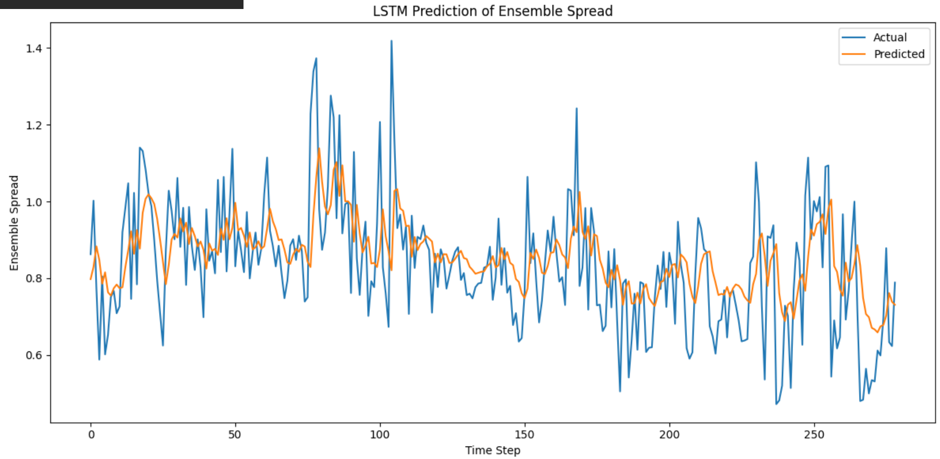


```
model = Sequential()
model.add(LSTM(50, activation='relu', input_shape=(X_train.shape[1], X_train.shape[2])))
model.add(Dense(1))
model.compile(optimizer='adam', loss='mse')

# Train model
model.fit(X_train, y_train, epochs=20, batch_size=32, validation_data=(X_test, y_test))
```



Training and Validation Loss over 50 Epochs

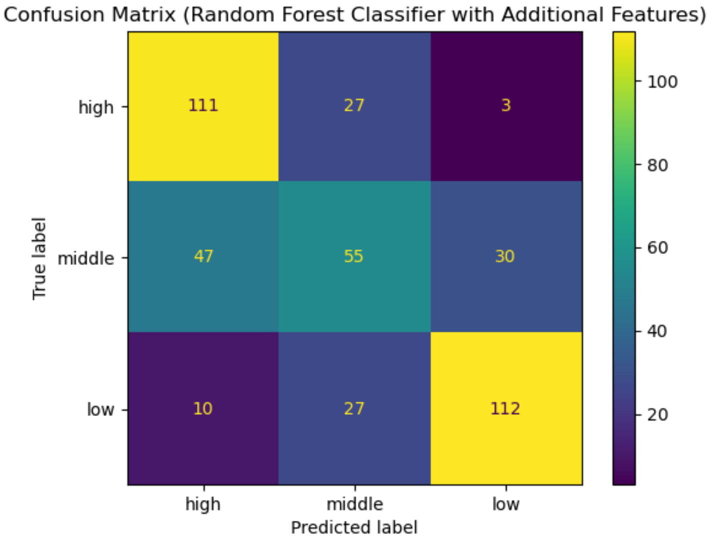


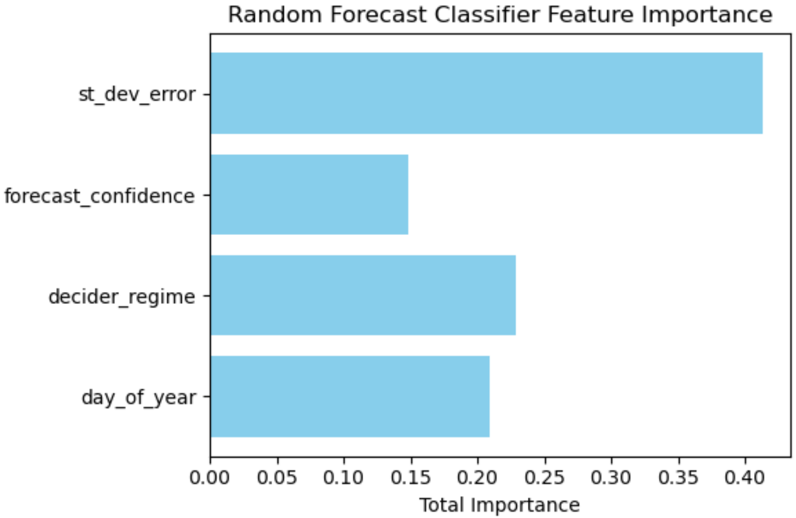
LSTM Prediction of Ensemble Spread

Results

Random forest classifier

Features: decider regimes, forecast confidence, day of the year, standard deviation of the error /
Target: ensemble spread class





Model accuracy score ~0.66

Conclusions and *future work*

- I tried different ML methods to experiment and getting more confidence and to improve the prediction.

- Generally, decider regimes and forecast confidence did not show a good predictive power for ensemble spread (RMSE similar, not shown). Instead, the standard deviation of error (calculated on the control forecast only) was quite important as a feature, as it improved both LSTM and RF classes predictions.

- One possible limitation is that there was not clear relationship between inputs and outputs originally, so the challenge was to exploit the data in order to find such relationship. Also, it would be good to have more past weather forecast to train on.

- *Try other ML methods: I guess that just linearly regressing on the standard deviation of error could yield good enough results.*

Tack!

Any questions ?

E-mail:
carlo.cafaro@metoffice.gov.uk