Prediction Methods for Numerical Weather Model Uncertainty

Patrik Durdevic, Megi Jaupi, Ryan Lucas, Thomas Wright

November 2022

1 Introduction

Few topics dominate day to day life on the same scale as weather. From natural disaster management to deciding whether or not to carry an umbrella, everyone is influenced and affected by the weather each and every day. Weather also has great influence on energy markets, as weather forecasts have a direct impact on the supply and demand of energy. On the supply side, renewables are fully dependent on weather conditions. Additionally, a cold wave, for example, can increase the demand used for heating purposes. One can think of a plethora of such examples that can drive namely every contract used in energy markets. Yet, regardless of the great amount of resources being used for the task, producing high quality granular weather forecasts remains extremely difficult. This introduces large amounts of uncertainty in the forecasts produced, which becomes in and of itself an important and challenging prediction task.

Uncertainty in weather forecasts comes from various sources. Firstly, it is ubiquitously known that weather systems are chaotic in nature. Additionally, no model can capture the full complexity of the physical world. Finally, whether we are dealing with physical or data-driven models, there may be errors in the data we have collected. All this considered, it is not enough to provide point estimations of weather forecasts, but rather a reliable estimation of the probability distribution of weather parameters. To characterize this probability distribution and to quantify the uncertainty that comes with trying to estimate it, weather forecast institutes run an ensemble of weather prediction models in parallel, subsequently generating an uncertainty space that can be measured.

With an increase in the forecast horizon, the uncertainty measured by the ensemble models increases exponentially Figure 1. Improving our capacity to understand and predict the uncertainty of numerical forecasts can have wide reaching impact. In the case of extreme weather events, for instance, estimating the severity of the event, i.e. the range of the values the weather parameter may take, is important for policy makers to understand the best course of action for evacuation or protective protocols. On the other spectrum of use cases stands the one that guides the methodology of this paper; the uncertainty and variability of weather forecasts drive the volatility of energy contracts, impacting trading strategies and positions. In this paper, in collaboration with Balyasny Asset Management (BAM)'s energy trading desk, we undertake the task of forecasting the uncertainty generated by physical ensemble models, using statistical and machine learning models.

2 Data

The available dataset comprises of two classes of physical weather models, the Global Ensemble Forecast System (GEFS) and European Centre for Medium Range Weather Forecasts (EC15). We make use of the data for two weather indicators: wind and temperature. Wind data is present for 4 European countries, namely France, Netherlands, Great Britain and Germany for EC15 and for the last two countries for GEFS. Temperature data is present for the same 4 countries for GEFS, whereas for EC15, only the France and Germany data for temperature are considered.

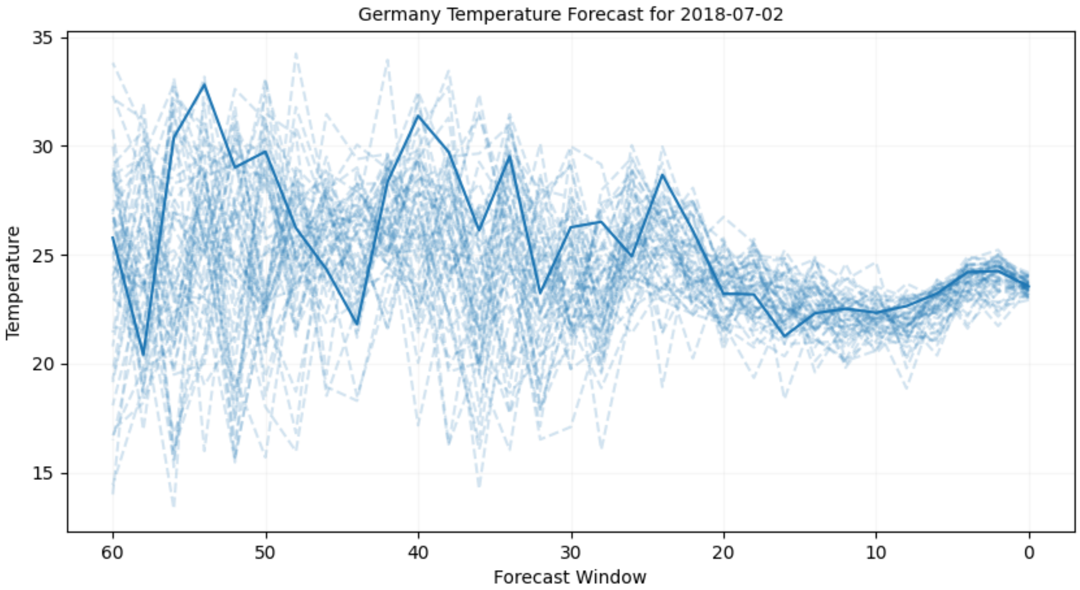

Figure 1: Depicting EC15 Temperature forecasts for Germany on 2018-07-02. Every dashed line represents the forecasts of an ensemble. The solid line represents the forecasts of Model_0. Clearly, as the forecast window decreases, i.e. we approach the date of interest, the models converge towards similar forecasted values. The dashed blue lines are a graphical representation of the uncertainty space generated by the physical models.

EC15 models produce their forecasts twice a day, at midnight and noon respectively. Each time the model runs, it predicts on a time window of 6h for the next 60 time windows. GEFS models provide more granular forecasts. They run 4 times a day, predicting again on a time window of 6h for the next 60 time windows. EC15 data is available from the 29th of March 2011 until the 28th of February 2022. Similarly, EGFS data is available from the 22nd of February 2011 until the 28th of February 2022.

Both physical model classes contain a Model_0 representing the best forecast and containing the physical model formulation. This model formulation is kept constant, but the initial conditions of Model_0 are slightly perturbed to generate each Ensemble_i prediction. Each physical model class, per country, per weather indicator contains a number of Ensemble predictions that behave similarly to Model_0, i.e at each forecasting point they output a predicted value for the next 6h, for the next 60 time windows. As previously mentioned, a slight perturbation in the initial conditions can result in very different forecast values for the weather indicators, due to their dynamic nature. This set of forecasts for every time window characterizes the uncertainty space generated by the physical models.

3 Methods

3.1 Overall approach

We approach the problem of uncertainty quantification and prediction from the perspective of predicting standard deviation among the ensemble members. We take this standard deviation as a proxy for the level of disagreement among the models, and hence an indication of how uncertain these models are about the target outcome. Through conversations with BAM, they communicated that they tend to trade daily, hence this forecast horizon was of primary concern. The standard deviation of ensemble models making forecasts on a 24-hour horizon is therefore the main quantity of interest throughout our project.

Mathematically, denote the forecasts of ensemble models making k-step predictions as $\hat{y}_{t+k,i}$, where i is the ensemble member. Now denote $\sigma_{k,t} = \frac{\sum_{i=1}^{N}(\hat{y}_{t+k,i} - \bar{y}_{t+k})}{N}$ as the standard deviation among these predictions. Here $\bar{y}_{t+k}$ is the mean forecast of the ensemble members and N is the total number of forecasts made. The goal of our study is to predict $\sigma_{24,t}$ at a forecast-horizon of 24 hours. Hence we fit various forecasting models of the form:

$$\hat{\sigma}_{24,t+24} = f(X_t)$$

where X_t, as we discuss, is comprised of information relating to the standard deviation of longer-horizon models, as well as the forecasts made across multiple geographies. For f, the forecasting model, we trialed simple ARIMA and ARIMAX models, as well as Sparse Regression, Vector Autoregression and LightGBM.

3.2 Leveraging multi-horizon forecasts

Our modelling approach revolves around using the standard deviation of longer-horizon models to predict the standard deviation of short-horizon models (namely 24 hour models). That is, when we seek to predict $\sigma_{24,t+24}$, which in principle comprises of models making predictions for $t + 48$, we leverage the known standard deviation of models making predictions on a 48-hour forecast horizon ($\sigma_{48,t}$). Note that there is no look-ahead bias here, since $\sigma_{48,t}$ and $\sigma_{72,t}$ are known at time t.

A natural objection to this modelling approach is: ”why not just use $\sigma_{48,t}$ as the prediction $\hat{\sigma}_{24,t+24}$, since both values relate to the same terminal date?”. Notice however that $\sigma_{48,t}$ is the standard deviation of models making forecasts on a 48 hour horizon. This quantity is naturally higher than a prediction for the standard deviation of models making 24 hour predictions. A visual intuition for this phenomenon is displayed in Figure 2. σ_{48} and σ_{72} are generally on a higher scale than σ_{24}, and hence cannot be compared directly. However, it is also clear that the quantities are highly correlated.

It clearly appears to be a promising direction to infer $\sigma_{24,t+24}$ using the value of $\sigma_{48,t}$ or $\sigma_{72,t}$. In Figure 3, we postulate a mechanism by which this relationship is predictive. Here $\sigma_{48,t}$ is the standard deviation 48 hours ahead. Notice that this standard deviation ’includes’ that of the shorter horizon, hence capturing possible developments in uncertainty that occur between $t + 24$ and $t + 48$. Since the physical models have more of a window into the future than autoregressive trends, including this standard deviation should capture more than an autoregressive relationship with σ_{24} might.

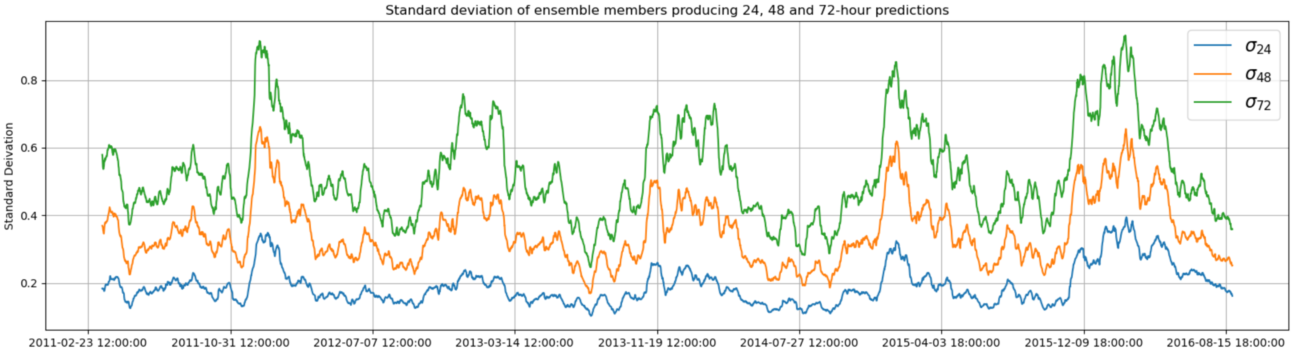

Figure 2: σ_{24}, σ_{48} and σ_{72} are fundamentally different quantities. Predicting one directly using the other does not make sense. However, they are highly related and hence the value of one can be used to inform the forecast for another.

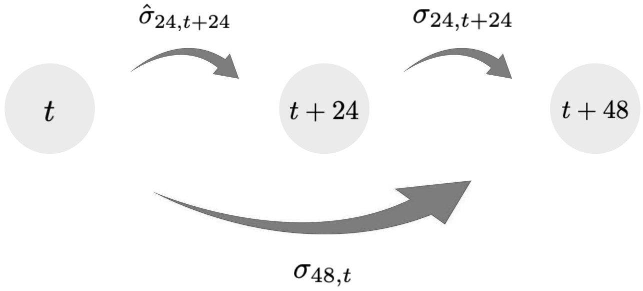

Figure 3: Leveraging standard deviation of forecasts made on longer horizons to inform shorter-term predictions. Here $\sigma_{48,t}$, which is the standard deviation of models making forecasts on a 48-hour forecast horizon, is known at time t. This standard deviation can be used to inform what the standard deviation of 1-day forecasting models will be in 24 hours.

3.3 Autogressive Integrated Moving Average Models (ARIMA/ARIMAX)

The Auto-Regressive Integrated Moving Average (ARIMA) is a linear model for forecasting univariate time series in which the predictors consist of previous values of the dependent variable and/or of the forecast errors, called lags. The lags of the stationary series are called ”autoregressive” terms, controlled by the p parameter, whereas the lags of the forecast errors are called ”moving average” terms controlled by the q parameter. ARIMA(p,d,q) is a general class of models for time series forecasting that allows differencing the series at order d to make it stationary. The transformed series is an ”integrated” version of the stationary series. The general ARIMA(p,d,q) model of the differenced time series y_t and error terms ϵ_t is fitted as follows:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + ... + \theta_q \epsilon_{t-q}$$

To apply ARIMA for our use case, we extend it to the ARIMAX model, adding a linear interaction of external variables. We also use an autoregressive order of 2 and ignore the moving average terms, yielding to the following equation:

$$\sigma_{24,t} = c + \beta_1 \sigma_{48,t} + \beta_2 \sigma_{72,t} + \phi_1 \sigma_{24,t-k} + \phi_2 \sigma_{24,t-2k}$$

This produces an iterative forecast for the quantity that we want to estimate:

$$\hat{\sigma}_{24,t+24} = c + \beta_1 \sigma_{48,t+24} + \beta_2 \sigma_{72,t+24} + \phi_1 \hat{\sigma}_{24,t+24-k} + \phi_2 \hat{\sigma}_{24,t+24-2k}$$

where $k = 6$ for GEFS and $k = 12$ for EC15.

3.4 Vector Autoregression

A Vector Autoregression (VAR) is a time series forecasting model that allows you to model several variables that evolve through time. In VAR, we estimate linear coefficients both on the variables themselves (autoregressive relationships) and also between variables in the system. A generic form VAR(2) with two variables would be estimated as follows:

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + \begin{bmatrix} \phi_{1,1} & \phi_{1,2} \\ \phi_{2,1} & \phi_{2,2} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} \gamma_{1,1} & \gamma_{1,2} \\ \gamma_{2,1} & \gamma_{2,2} \end{bmatrix} \begin{bmatrix} y_{1,t-2} \\ y_{2,t-2} \end{bmatrix} + \begin{bmatrix} e_{1,t} \\ e_{2,t} \end{bmatrix} \tag{1}$$

The useful aspect of VAR, above using separate ARIMAX models, is its ability to perform iterative forecasting. Once one-step forecasts are made, these forecasts are 'fed-forward' to predict $t+2$ and so on, as follows:

$$\begin{bmatrix} \hat{y}_{1,t+1} \\ \hat{y}_{2,t+1} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + \begin{bmatrix} \phi_{1,1} & \phi_{1,2} \\ \phi_{2,1} & \phi_{2,2} \end{bmatrix} \begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} + \begin{bmatrix} \gamma_{1,1} & \gamma_{1,2} \\ \gamma_{2,1} & \gamma_{2,2} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} \tag{2}$$

$$\implies \begin{bmatrix} \hat{y}_{1,t+2} \\ \hat{y}_{2,t+2} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + \begin{bmatrix} \phi_{1,1} & \phi_{1,2} \\ \phi_{2,1} & \phi_{2,2} \end{bmatrix} \begin{bmatrix} \hat{y}_{1,t+1} \\ \hat{y}_{2,t+1} \end{bmatrix} + \begin{bmatrix} \gamma_{1,1} & \gamma_{1,2} \\ \gamma_{2,1} & \gamma_{2,2} \end{bmatrix} \begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} \tag{3}$$

Since our data is available on a 6-hour and 12-hour frequency (for GEFS and EC-15 respectively), this allowed us to make predictions for 24-hours ahead, since we always predict 6 or 12 hours ahead and use these predictions to feed forward towards a 24 hour prediction. Thus our VAR model for GEFS standard deviation was specified as follows:

$$\begin{bmatrix} \sigma_{24,t} \\ \sigma_{48,t} \\ \sigma_{72,t} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} + \begin{bmatrix} \phi_{1,1} & \phi_{1,2} & \phi_{1,3} \\ \phi_{2,1} & \phi_{2,2} & \phi_{2,3} \\ \phi_{3,1} & \phi_{3,2} & \phi_{3,3} \end{bmatrix} \begin{bmatrix} \sigma_{24,t-6} \\ \sigma_{48,t-6} \\ \sigma_{72,t-6} \end{bmatrix} +, ..., + \begin{bmatrix} \gamma_{1,1} & \gamma_{1,2} & \gamma_{1,3} \\ \gamma_{2,1} & \gamma_{2,2} & \gamma_{2,3} \\ \gamma_{3,1} & \gamma_{3,2} & \gamma_{3,3} \end{bmatrix} \begin{bmatrix} \sigma_{24,t-24} \\ \sigma_{48,t-24} \\ \sigma_{72,t-24} \end{bmatrix} \tag{4}$$

We then use this higher frequency information to make an iterative forecast for $t+24$:

$$\begin{bmatrix} \hat{\sigma}_{24,t+24} \\ \hat{\sigma}_{48,t+24} \\ \hat{\sigma}_{72,t+24} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} + \begin{bmatrix} \phi_{1,1} & \phi_{1,2} & \phi_{1,3} \\ \phi_{2,1} & \phi_{2,2} & \phi_{2,3} \\ \phi_{3,1} & \phi_{3,2} & \phi_{3,3} \end{bmatrix} \begin{bmatrix} \hat{\sigma}_{24,t+18} \\ \hat{\sigma}_{48,t+18} \\ \hat{\sigma}_{72,t+18} \end{bmatrix} +, ..., + \begin{bmatrix} \gamma_{1,1} & \gamma_{1,2} & \gamma_{1,3} \\ \gamma_{2,1} & \gamma_{2,2} & \gamma_{2,3} \\ \gamma_{3,1} & \gamma_{3,2} & \gamma_{3,3} \end{bmatrix} \begin{bmatrix} \sigma_{24,t} \\ \sigma_{48,t} \\ \sigma_{72,t} \end{bmatrix} \tag{5}$$

3.5 Sparse Regression

In data settings with a large number of predictors relative to the number of observations, dimensionality reduction is often required ahead of the modelling step. While there are many different approaches to dimensionality reduction, sparse regression is one that maintains the highest level of interpretability while remaining performant.

Sparse regression relies on selecting the best subset of features available, namely, selecting $k << p$ features. For this study, the cross correlation between the features and the target was used to pick the k most important covariates. The problem then becomes a simple linear regression task on the smaller dataset:

$$\hat{y} = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k \qquad \text{s.t. } X_i \text{ is column } i \text{ of } X$$

Applied to our setting, an example of what we could see with $k = 3$ is:

$$\hat{\sigma}^{DEU}_{24,t+24} = \beta_0 + \beta_1 \sigma^{DEU}_{48,t} + \beta_2 \sigma^{FRA}_{48,t} + \beta_3 (\sigma^{DEU}_{72,t} - \sigma^{NLD}_{72,t})$$

3.6 LightGBM

Boosted trees have gained lots of popularity due their impressive general performance, at the cost of a high level of interpretability. For this project, we leverage LightGBM, one of the most popular methods in the space of Boosted trees today.

Gradient boosted decision trees in general work by training a weak initial model and checking the performance on the dataset. The prediction error gives weights to the training set, emphasizing examples which were predicted incorrectly. Then, a new model is trained with a goal of fixing these errors, forming an ensemble. Subsequent trained models form an ensemble which makes predictions by aggregating those of all members. LightGBM is based on this algorithm, and differs from similar solutions by focusing on leaf-wise growth.

Indicator	Country	Model	ARIMAX	VAR	Sparse Regression (k=5)	LightGBM
Wind	FRA	EC15	0.06	0.63	0.66	0.66
Wind	NLD	EC15	0.02	0.60	0.61	0.59
Wind	GBR	EC15	-0.06	0.47	0.46	0.50
Wind	GBR	GEFS	0.17	0.64	0.64	0.43
Wind	DEU	EC15	-0.02	0.60	0.59	0.61
Wind	DEU	GEFS	0.15	0.66	0.64	0.41
Temp.	FRA	EC15	-0.02	0.47	0.45	0.47
Temp.	FRA	GEFS	0.16	0.43	0.42	0.14
Temp.	NLD	GEFS	0.16	0.47	0.45	0.18
Temp.	GBR	GEFS	0.18	0.63	0.53	0.20
Temp.	DEU	EC15	-0.06	0.55	0.51	0.53
Temp.	DEU	GEFS	0.18	0.47	0.41	0.20

Table 1: Validation set R^2 performance for $\sigma_{24,t+24}$ being used for model selection

4 Experiments & Results

We split the dataset in a 50-30-20 ratio for the training, validation and test set respectively, preserving the time dimension of the series. This results in a validation set of 2393 data points for EC15, 4827 data points for GEFS and in a test set of 1597 data points for EC15 and 3220 for GEFS. For the time series models, the augmented Dickey–Fuller test was conducted for each quantity of interest. All tests were positive for stationarity. This was expected, considering that our quantities of interest are standard deviations. Considering this fact, we did not perform differencing in the time series. We use the validation set to evaluate the predictive ability of the models and for model selection. Then, we use the test set to evaluate the impact of our model in improving the uncertainty space quantification.

4.1 Model performance

In Table 1, we summarize the performance of the models on the validation set for $\sigma_{24,t}$ as measured by R^2. First notice that the classical approach of time series forecast, namely ARIMA, and in our case the enhanced version of the model with external predictors as explained in Section 3.3 performs really poorly. We conducted many experiments to see weather introducing moving average terms and longer autoregressive lags would improve the model, but to no avail. This was an indicator that a simple model could not capture the complexity of the weather uncertainty.

Multivariate and more complex models perform much better than our initial attempt, with VAR seemingly performing better on the validation set overall. VAR and Sparse regression perform similarly for both GEFS and EC15, while VAR has a slight edge over the Sparse Regression in most cases. A surprising result is that VAR has a significant edge over LightGBM on the GEFS models. R^2 scores for LightGBM show us that the tree-based model has an especially hard time modelling GEFS model behavior. For the EC15 model, it manages to produce results comparable to VAR and Sparse Regression.

We decided to choose our model based solely on the performance on the validation set, however being aware of the importance of model interpretability, we devote the whole next section to it. Observing that the VAR model performs best on the validation set, we choose it as the model to evaluate the impact of our forecasts. For the fitted model, we compute the performance of VAR model for predicting both $\sigma_{24,t+24}$ and $Model_{0_{24,t+24}}$ on the test set with results as in Table 2. Comparing the two, we can see that scores on the test set are lower than on the validation set.

Indicator	Country	Model	$\sigma_{24,t+24}$	$Model_{0_{24,t}}$
Wind	FRA	EC15	0.68	0.97
Wind	NLD	EC15	0.61	0.94
Wind	GBR	EC15	0.57	0.994
Wind	GBR	GEFS	0.72	0.94
Wind	DEU	EC15	0.69	0.96
Wind	DEU	GEFS	0.69	0.94
Temp.	FRA	EC15	0.68	0.996
Temp.	FRA	GEFS	0.68	0.995
Temp.	NLD	GEFS	0.62	0.99
Temp.	GBR	GEFS	0.63	0.993
Temp.	DEU	EC15	0.57	0.996
Temp.	DEU	GEFS	0.70	0.995

Table 2: Test set VAR R^2 performance for $\sigma_{24,t+24}$ and $Model_{0_{24,t+24}}$

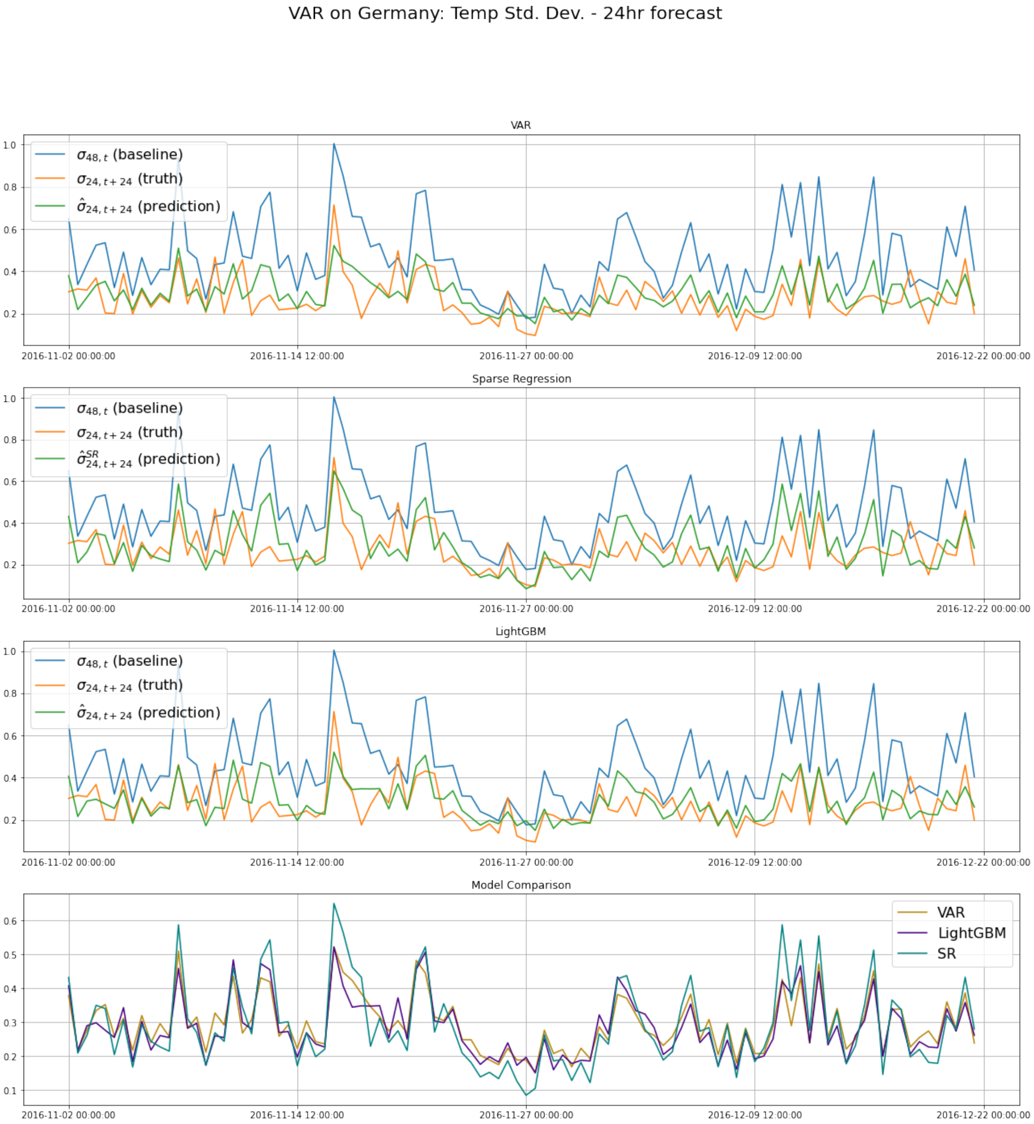

Figure 4: Ahead of predictive modelling, our knowledge of physical model uncertainty, and therefore weather uncertainty was limited to forecasts at time t: $\sigma_{48,t}$, our uncertainty from 48 hours out. Our prediction $\hat{\sigma}_{24,t+24}$ uses only knowledge available at time t and represents an understanding of the uncertainty from $t+24$, narrowing the uncertainty by over 30 %.

4.2 Model Interpretability

4.2.1 VAR

A VAR models interactions between several variables that are each treated as endogenous. Its ϕ coefficients can be interpreted in the same way as a regular linear regression with a single target. In particular, for two variables y and x, $\phi_{y,x}$ can be interpreted as "A unit increase in x leads to a $\phi_{y,x}$ increase in y". As an example, Figure 5 shows VAR coefficients for a model on wind in Great Britain from EC15. Notice that the constant coefficients on $\sigma_{24,t}, \sigma_{48,t}$ and $\sigma_{72,t}$ are naturally increasing, reflecting the natural increase in standard deviation as the forecast horizon increases. The interaction terms are also relevant. Notice that the autoregressive interactions are generally strong, but strongest of all are interactions that occur between long and short-horizon standard deviations relating to the same target. This strongly supports our earlier hypothesis, that longer horizon information can be leveraged early to uncover information about developments in uncertainty over the next 24 hours.

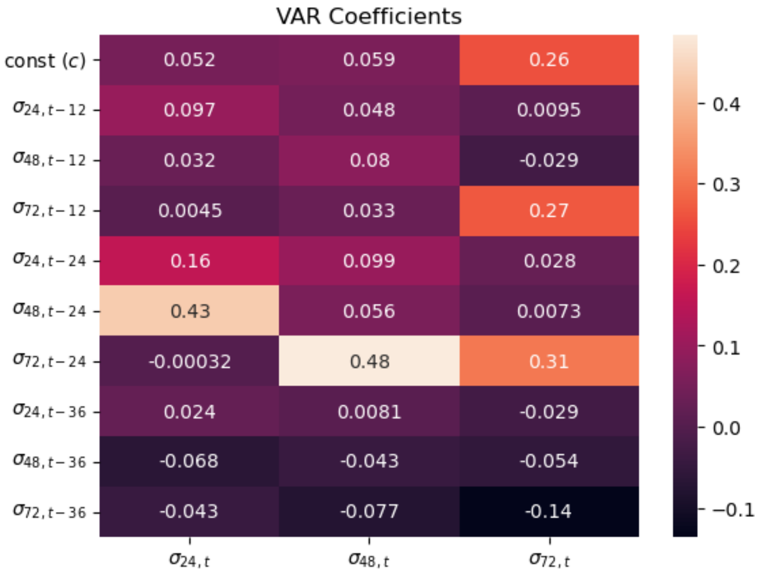

Figure 5: Coefficients of interaction for an EC15 GBR-Wind VAR model. Notice that the strongest interactions occur between long and short-horizon standard deviations relating to the same target. For instance, $\sigma_{24,t}$ and $\sigma_{48,t-24}$ interact strongly, while $\sigma_{48,t-24}$ is available earlier than $\sigma_{24,t}$ and hence can be used as a strong early indicator. This reaffirms the predictive relationship we hypothesized in subsection 3.2.

4.2.2 Sparse Regression

An important advantage to sparse regression over other dimensionality reduction techniques is its interpretability. Given a large number of features, we can identify which are most important for predicting our target. In the context of this project, we take advantage of this fact in order to derive insights in to correlations and trends across geographies and first order differences.

For example, when predicting $\sigma_{24,t+24}^{GBR,Wind}$, the resulting model is:

$$\begin{aligned}\hat{\sigma}_{24,t+24}^{GBR,Wind} &= 0.094 + 0.415\sigma_{48,t}^{GBR,Wind} + 0.051(\sigma_{48,t}^{GBR,Wind} - \sigma_{48,t}^{FRA,Wind}) \\ &+ 0.010(\sigma_{48,t}^{DEU,Wind} - \sigma_{48,t}^{GBR,Wind}) - 0.036(\sigma_{48,t}^{GBR,Temp} - \sigma_{48,t}^{FRA,Temp}) + 0.003(\sigma_{48,t}^{DEU,Temp} - \sigma_{48,t}^{GBR,Temp})\end{aligned}$$

Intuitively, the greatest influence is the 48-hour forecast uncertainty at time t, as the terminal time of this matches that of the terminal day of our target, namely, $t + 48$. This is followed by geographically differenced wind, GBR minus FRA and DEU minus GBR. Finally, we see two geographically differenced temperature features; GBR minus FRA and DEU minus GBR. These results indicate a link between wind uncertainty and temperature uncertainty, as well as geographic trends.

4.2.3 LightGBM

Since LightGBM models consist of numerous decision trees, we are not provided with any inherent or robust interpretation of how they make decisions. For that reason, Shapley values, originally developed for game theory, provide us a mechanism to infer which features contribute the most to making predictions.

Figure 6 shows that the most important feature (in Shapley interpretation of importance) for predicting $\sigma^{GBR,Wind}_{24,t+24}$ is $\sigma^{GBR,Wind}_{48,t}$. We see contributions from the difference between Germany's and Great Britain's 2-day wind forecasts followed by the difference between France's and Great Britain's 2-day wind forecasts. Finally, we have other features making small-scale contributions to this model's predictions.

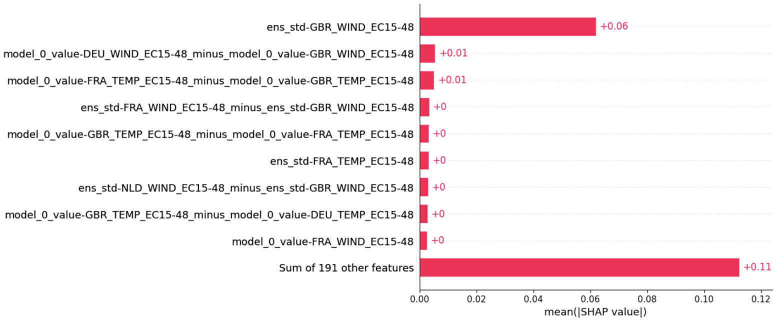

Figure 6: Feature importance for a LightGBM model predicting $\sigma^{GBR,Wind}_{24,t+24}$ using Shapley values.

Notably, the three most important variables identified by the tree based LightGBM and sparse regression match, even though they use them in completely different ways. Further analysis on geographic and temporal relations within the data is recommended based on these initial results on the interpretability of emerging trends and relationships.

5 Impact

In the interest of quantifying the impact of our models on downstream business tasks, we focus on two fundamental uncertainty metrics: coverage and spread. Coverage tells us if the true value falls within our predicted interval, while the spread tells us how wide out uncertainty band is. One can recognize the trade-off between the two metrics – if we increase the spread of our predictions, the coverage will be higher, and vice versa. An optimal prediction would have full coverage while and a narrow spread. We use an adjusted Winkler score as a measure of how good our forecasts are:

$$W_{\alpha,t} = \begin{cases} (u_{\alpha,t} - \ell_{\alpha,t}) + \frac{2}{\alpha}(\ell_{\alpha,t} - y_t) & \text{if } y_t < \ell_{\alpha,t} \\ (u_{\alpha,t} - \ell_{\alpha,t}) & \text{if } \ell_{\alpha,t} \leq y_t \leq u_{\alpha,t} \\ (u_{\alpha,t} - \ell_{\alpha,t}) + \frac{2}{\alpha}(y_t - u_{\alpha,t}) & \text{if } y_t > u_{\alpha,t}. \end{cases} \tag{6}$$

To understand the business impact of our modelling, we need to remember that the standard deviation of ensemble models making forecasts on a 24-hour horizon is the main quantity of interest throughout our project. On a specific day, physical models give the business sense of what the weather is going to be in 48 hours, while our models make predictions on what the 24 hour forecast is going to be the day after – both predicting for the same

Indicator	Country	Model	Improvement (%)
Wind	FRA	EC15	30.93
Wind	NLD	EC15	33.58
Wind	GBR	EC15	34.67
Wind	GBR	GEFS	34.69
Wind	DEU	EC15	30.14
Wind	DEU	GEFS	51.69
Temp.	FRA	EC15	28.31
Temp.	FRA	GEFS	49.89
Temp.	NLD	GEFS	50.89
Temp.	GBR	GEFS	53.76
Temp.	DEU	EC15	30.27
Temp.	DEU	GEFS	52.33

Table 3: Test set Winkler score improvement for a 24-hour forecast prediction over the initial 48-hour forecast

point in time. Therein lies the impact of our project – how much better is our prediction compared to the initial 48-hour forecast.

Having the Winkler score metric to measure just that, we compare scores for a 48-hour forecast with the 24-hour forecast predictions using VAR as a model of choice due to its performance against other models. Table 3 reports Winkler score improvements for different targets on the test set.

6 Conclusion and extensions

While this work achieved what it sought to, it also uncovered fruitful directions to extend and improve upon what was discovered. Performance-wise, focusing on methods not applied in this study could boost results, or indicate which methods are not suitable. In consideration of interpretability and insights, the work done in this project indicates that further analysis into geographic and temporal relationships could end up being incredibly meaningful. Lastly, we focused on predictions for $\sigma_{24,t+24}$ as it could prove to be a valuable quantity for energy trading. That said, examining performance at different time horizons and prediction intervals would be a valuable extension, as the value of the forecast to different domains shifts with the time horizon and prediction interval.

The work done throughout this project shows that it is possible to forecast weather uncertainty at a future date. Although we had to limit this paper's scope to a methodology that was feasible to implement in the limited time that we could allocate to this class, it is also important to mention other approaches and ideas that we experimented with. One such direction is extending the modelling from predictive methods to generative methods with Deep Generative Neural Networks. This extension is natural since rather than predicting a future interval of the uncertainty space, it acts as an ensemble substitute by learning the underlying distribution that the physical models induce.

Given the wide-reaching influence weather has on human and economic problems, the results of this work have the potential to enable better decision-making across industries. Policymakers can leverage this information to help improve natural disaster management. Furthermore, these predictions can help decision-makers and algorithms hedge against risk by scaling weather-dependent portfolios accordingly. Across a wide range of domains, there is no shortage of benefits that can be drawn from improved weather uncertainty knowledge.

7 Appendix

7.1 Error correction

Residuals are observed to be linearly correlated with the target, indicating that we are not catching the full variance of the underlying data. A linear model was trained on the residuals of the training data, as an error correction model. The final model is then:

$$\hat{f}^* = \hat{f} + \hat{f}_{error}$$

This model structure was tested on the sparse regression and successfully corrected the linear relationship between target value and residual. There is no obvious improvement or compromise with regards to performance, as the corrected model outperforms the original model roughly half the time.

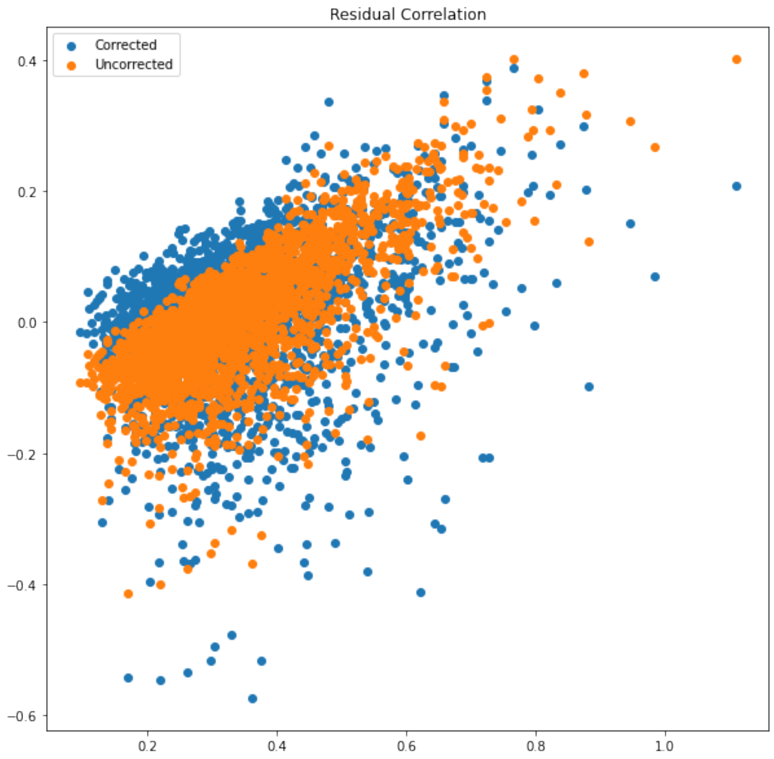

Figure 7: The corrected model residuals clearly has more random spread than that of the uncorrected model