Combined Statistical-Dynamical Downscaling of Wind Components over the British Isles and Surrounding Water

Ian Goddard

Master of Science

Data Science

School of Informatics

University of Edinburgh

2019

Abstract

Wind energy expansion in the United Kingdom has increased interest in solving challenges posed by large scale integration of intermittent power sources onto the electrical grid. To better understand the potential for wind power as an energy source, and answer key questions in relation to these challenges, a detailed understanding of the spatial and temporal variability of the UK wind resource is essential. Current methods for predicting wind resource across the UK, derived from data produced by *physically* based global atmospheric models, are limited to 30 km resolution due to their computational complexity. This thesis develops and evaluates several *statistical* methods for inferring 3 km wind data from 30 km data produced by a global atmospheric model. We propose several linear regression methods, specifically lasso and ridge regression to investigate the importance of the model covariates, and to achieve high prediction performance. We show a clear benefit of these regression methods compared to a commonly used, quick method for increasing spatial resolution, known as bi-linear interpolation. We provide a detailed discussion on the seasonal and spatial errors of the proposed models, and show that the statistical methods applied allow us to reduce the total relative errors by up to 15% compared to the baseline model. We conclude that these statistical methods allow us to reproduce high resolution reanalysis data with far lower computational expense than atmospheric models.

Acknowledgements

First, I would like to thank Dr Michael Guttman for his invaluable feedback and guidance during this project, his insights will stay with me long after this work. Secondly, I would like to thank Professor Gareth Harrison for providing the data necessary to conduct this work. Finally, I would like to thank my family for their continuous support throughout this degree.

Table of Contents

Chapter 1

Introduction

1.1 Motivation

In recent years, the growth of renewable energy technologies has drawn much attention to the challenges of large scale integration of intermittent power sources onto the electrical grid. In particular, due to the stochasticity of wind, there is inherent uncertainty about the security of supply [1], requiring quickly dispatchable electricity generation resources be kept on reserve to cope with sudden mismatch in supply and demand. Such dispatchable energy sources are costly and, at present, mostly fossil fuel based due to their quick dispatch times [2, 3]. Hence, it is economically and environmentally beneficial to reduce uncertainties associated with all intermittent power sources, including wind power. Whilst there has been a sharp growth in the wind power industry worldwide [4], wind power is of particular importance in the United Kingdom, and will be a key driver in reaching the UK's ambitious energy targets [5]. Gaining a detailed understanding of the spatial and temporal variability of the total energy available from the wind, known as the wind resource [1], will enable accurate assessments of the potential for wind power to meet these targets.

To obtain robust assessments of wind resource, we require a dataset which satisfies two criteria. Firstly, sufficient temporal extent to capture the average temporal variability, known as a *climate normal* for the wind. This period is defined to be a three decade average by the National Oceanic and Atmospheric Association (NOAA)[1] [7]. Secondly, we require this dataset have high and regular spatial resolution to capture spatial variability on local scales (~ 5 km), such that local scale topographic features,

[1] A thirty year period for defining climate normals is convention, arising from the fact that there only existed thirty years worth of data, at the time that climate normals were first calculated [6].

such as mountain ranges and valleys which influence surface flows, are resolved within the dataset [8].

Previous estimates of the UK wind resource have been conducted through analysis of recorded wind speeds at wind measurement stations across the UK. However, a clear limitation of this approach is that obtaining accurate assessments is only possible close to measurement stations. A large network of these stations exists within the UK, however, their spacing is only guaranteed to be less than 50km [9], which is too sparse to resolve local scale features. Thus, this approach is not able to provide the spatial resolution we require for a full accurate assessment across the whole of the United Kingdom.

Alternatively, *reanalysis* data can be used for wind resource assessment. Reanalysis data is produced through atmospheric modelling, constrained by assimilation of historic measured meteorological data, to produce historic time series for many meteorological variables at uniform resolution. As these datasets are produced using historic data, they often extend well beyond the thirty year period we require to obtain a climate normal. Notable state-of-the-art examples are the European Centre for Medium Range Weather Forecasts (ECMWF) ERA5 reanalysis dataset [10], available from 1979 to present, and NASA's Modern-Era Retrospective analysis for Research and Applications 2 (MERRA 2) [11], providing data from 1980 to present. Whilst both these datasets satisfy the temporal and regularity criteria, their spatial resolutions are (~ 30 km $\times \sim 30$ km) and (~ 50 km $\times \sim 50$ km), respectively. The spatial resolution is limited due to the large computational costs ($\sim 100{,}000$ of processor hours [12, 13]) associated with global atmospheric modelling. Simulations by ECMWF show that increases in model resolution correspond to greater than exponential increases in the number of required computer cores (Figure 1.1), predicting that global atmospheric models with the spatial resolution we require will not be available for several years.

To obtain a dataset which satisfies all criteria imposed, it is possible to *downscale*[2] output from global atmospheric models to increase their spatial resolution. In this thesis we explore several methods to downscale data for wind components for the British Isles and surrounding water.

[2]Downscaling, not to be confused with downsampling, is the term used in geosciences to mean increasing spatial resolution.

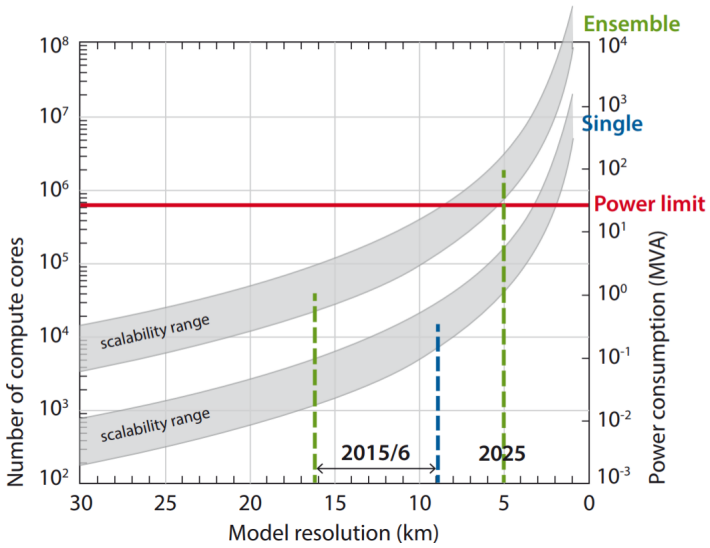


Figure 1.1: *Shows simulations carried out by ECMWF for a range of models with hypothetical model resolutions, note the logarithmic left and right axes. Figure provided by [14].*

1.2 Contribution

The contributions of this thesis are particularly aimed at addressing one of the main limitations of the dataset detailed and discussed in [12]. Hawkins et al.[12] have created an 11-year, high resolution $\sim$ (3 km $\times$ 3 km) reanalysis dataset for wind components over the British Isles and surrounding water. However, as stated in Section 1.1, a thirty year period is desirable as this is considered the industry standard to define the 'normal' temporal variability of a meteorological variable. The authors of [12] acknowledge the temporal extent of their reanalysis dataset as a limitation, motivating our exploration of methods to allow for the extension of this dataset.

Here, we build and evaluate several linear regression based methods to statistically downscale low resolution reanalysis data, such that we can accurately reproduce high resolution reanalysis data at the scale of the dataset produced in [12]. The full contributions of this thesis are as follows:

- We build several regression models that achieve higher accuracy than the commonly used method of bi-linear interpolation.
- We show that ridge regression allows us to achieve better performance over previously explored, unregularised multivariate linear regression.
- We find, through the use of interpretable linear regression methods, that there is

clear preference for atmospheric pressure variables to be included in the models, when predicting the u wind component. This indicates that in future studies, pressure fields should be included if one aims to predict the u component of wind.

- By estimation of the generalisation performance of the models, we show that our method offers a computationally inexpensive way to generate accurate, high resolution reanalysis data. However, further work is required to understand how our models perform at predicting real wind component time series.

Chapter 2

Background

2.1 Meteorological Background

To understand the steps taken in developing the models for this thesis, it is important to first provide an explanation of the factors which influence the quantities we wish to predict, namely the west to east (u) and south to north (v) wind components. Fundamentally, winds are created by the flow of air between regions of high and low pressure, often occurring due to non-zero temperature gradients. Once a flow begins, its motion is influenced by a mix of several forces: Large weather systems driven by large scale pressure gradients, land surface heterogeneity creates local scale pressure gradients and the Coriolis force, all of which act to produce turbulent and complex motion of near surface flows. [15].

In the statistical analysis conducted in this study, we include auxiliary variables, such as temperature and pressure, which may contain predictive information on the effects of these forces. Further, we include a meteorological variable known as *relative vorticity* which measures the vertical component of the angular velocity of flows relative to the Earth. This variable was included as a measure of the magnitude of vertical motion which may contain information on coupling between near-surface and upper atmospheric winds [16].

2.2 Global Atmospheric Models and Reanalysis Data

To model not only the dynamics of the wind, but the dynamics of the full atmospheric system, researchers have been developing atmospheric models since the early 20th century [17]. All atmospheric models are based on a set of non-linear partial differential

equations, known as *primitive equations*, which in almost all cases have no analytical solutions and must be solved using numerical integration methods [12]. These equations are based on physical interpretations of the atmosphere from global down to local scale. However, resolving atmospheric dynamics on all spatial scales is intractable and models are often restricted either by geographic domain over which they operate, or by their spatial resolution [12]. As discussed in the Section 1.1, state-of-the-art global atmospheric models are currently limited to $\sim 30 \times 30$ km resolution.

Reanalysis data is produced using atmospheric models constrained by historic measured values to create a hind-cast of the climate and obtain historic time series for many meteorological variables. Spatial resolution of reanalysis data is restricted to that of the atmospheric model which produced the data, meaning that to obtain high resolution reanalysis data from atmospheric models, we incur the high computational costs associated with atmospheric modelling.

2.2.1 Low Resolution Reanalysis Data

The recently produced ERA-5 dataset from (ECMWF) [10] is used in this study. This dataset is produced via reanalysis, using a global atmospheric model, and consists of hourly data from 1979 to present for many meteorological variables, including the u, v wind components. As this model operates over the globe, the spatial resolution is limited to $0.25^\circ \times 0.25^\circ$ longitude/latitude $\sim$(30 km $\times$ 30 km). This dataset offers benefits over other comparable datasets, such as MERRA2 [11], due to the higher spatial resolution of ERA5. Further, in particular for this work, previous literature has shown that ERA5 outperforms MERRA2 when used to predict near surface wind speeds [18].

2.2.2 High Resolution Reanalysis Data

The predictand data for this study comes from a regional climate model (RCM) covering the British Isles and surrounding water, and is extensively detailed and validated in [12]. This dataset has spatial resolution of roughly 3 km $\times$ 3 km, covers a time period from 2000 to 2011 with hourly data for the u, v wind components at three heights (10m, 80m and 100m). This dataset was produced using the High-end Computation Terrascale Resource (HECToR) [19] in monthly batches, each batch requiring 12 hours simulation time using 512 processors [12]. For the full 11 year period this equates to

$12 \times 12 \times 11 = 1584$ hours[1] (>2 months) of HECToR time or approximately 800k processor hours. A full 30-year period would require roughly 6 months of HECToR time, or 2.5M processor hours.

2.3 Downscaling Background

As mentioned in the Section 1.1, it is possible to downscale output from global atmospheric models, to increase spatial resolution. Techniques for downscaling low resolution data fall into two broad categories: *dynamical* and *statistical*.

Dynamical donwscaling (Figure 2.1 (red path)) tackles the problem using a physical approach through construction of an RCM spanning an area of hundreds to thousands of kilometres. Whilst this method has previously been shown to produce accurate results [20, 12, 21], the physical approach to downscaling brings large computational costs [22, 12, 23, 24] due to the expense of atmospheric modelling.

Alternatively, methods of statistical downscaling are based upon construction of a transfer function capable of transforming large scale data ($>\sim 20$ km) to local scales ($<\sim 5$ km) [25, 26] (Figure 2.1 (blue path)) . Common approaches have taken local scale data to be measured wind speeds [25, 26, 27]. A limitation of these studies however, is that the spatial coverage of the statistical method is limited to the regions covered by the measured data. For example, to obtain an accurate model for regions over the sea, this method would require long time series of offshore measured wind speed data, which are scarce [22].

Whilst both statistical and dynamical methods are often applied separately, they are in fact complementary (Figure 2.1 (red + yellow path)). By producing a small number of years of dynamically downscaled wind data, a transfer function is constructed to map low resolution reanalysis data to dynamically downscaled high resolution reanalysis data. This limits the computational effort spent on dynamical downscaling and is the approach considered in this thesis.

2.4 Methodology Background

Here we outline the statistical background required to understand the methodology used in this thesis; we introduce linear methods for regression and discuss the relative merits of the different approaches. A large portion of this material is based on [28].

[1] We calculate this using the formula: time taken per month $\times$ number of months $\times$ number of years

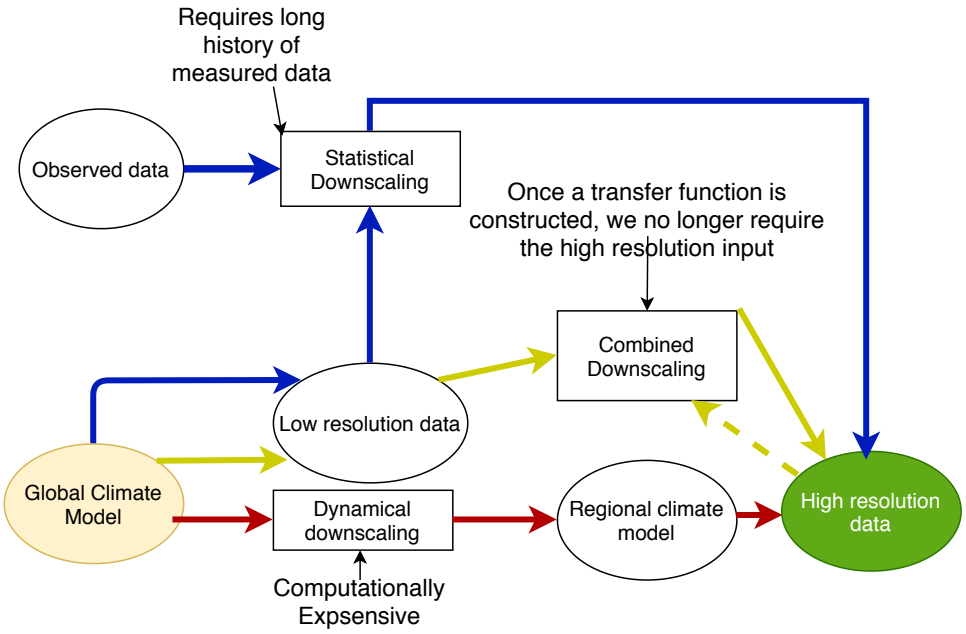


Figure 2.1: Schematic representation of the possible routes to obtain downscaled meteorological data. All paths begin with a global climate model (yellow) and end at high resolution data (green). The blue path shows the purely statistical approach, whilst the red path shows the dynamical approach. The yellow + red paths show the combined approach. By using this approach, we aim to eliminate the dotted yellow arrow, after learning a relationship between the low resolution and high resolution datasets, such that the high resolution data is no longer needed as input, to generate more high resolution data.

2.4.1 Methods for Linear Regression

Linear regression models are defined by the property that they are linear in the parameters of the model. Given the values of some p-dimensional input vector $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ of input variables, the simplest linear regression model takes a linear combination of these variables to predict one or more target variables y. The linear model is defined by (2.1), where the β_j's are the unknown model parameters and β_0 denotes the bias.

$$y = \beta_0 + \sum_{j=1}^{p} x_j \beta_j \tag{2.1}$$

To estimate the parameter values, $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_p]^\intercal$, we often have a set of n training samples and minimise the residual sum of squares (RSS) over all training

samples. This is given by (2.2), where $x_j^{(i)}$ indicates the j^{th} dimension of the i^{th} training input, $y^{(i)}$ denotes the i^{th} training target and $\boldsymbol{\beta}$ denotes the vector of model parameters.

$$RSS(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y^{(i)} - \beta_0 - \sum_{j=1}^{p} x_j^{(i)}\beta_j)^2 \tag{2.2}$$

If we write the set of n training samples as $\mathbf{X}$ where each row in $\mathbf{X}$ is a training sample $\mathbf{x}^{(i)} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_p^{(i)})$ with a 1 in the first column, and let $\mathbf{y}$ be the corresponding vector of target samples, we can write equation 2.2 in matrix form [28].

$$RSS(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{2.3}$$

Assuming that $\mathbf{X}$ has full rank, making $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ positive semi-definite and thus invertible, we can write the closed form solution for the optimal parameter set $\hat{\boldsymbol{\beta}}$.

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \tag{2.4}$$

2.4.2 Shrinkage Methods

A well known result in statistics, known as the Gauss-Markov theorem states that the least squares estimate of the parameter set, $\hat{\boldsymbol{\beta}}$, has the smallest variance among all linear *unbiased* estimates [28]. However, there exist many methods which result in biased estimates that can offer a lower mean square error, by trading a small amount of bias for a large reduction in the variance of the parameter estimates [28]. Several of these methods, known as shrinkage methods, obtain biased estimates by reducing the magnitude of the regression coefficients. Here we discuss two common shrinkage methods, namely ridge and lasso regression.

2.4.2.1 Ridge Regression

Ridge regression, more commonly known as L2 regularisation, penalises the sizes of the regression coefficients through an additional term in the cost function (2.5), where λ controls the strength of the parameter penalty.

$$\hat{\boldsymbol{\beta}}^{ridge} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left[\sum_{i=1}^{n} (y^{(i)} - \beta_0 - \sum_{j=1}^{p} x_j^{(i)} \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right] \tag{2.5}$$

A useful property of ridge regression is the ability to deal with highly correlated inputs. In a general linear regression model, correlated variables may lead to ill-determined coefficients, as a large positive coefficient may cancel out with a large negative coefficient, if the variables associated with those coefficients are highly correlated. By imposing a penalty on the magnitude of the coefficients, ridge regression weakens the effect of correlated inputs.

2.4.2.2 Lasso regression

As with ridge regression, lasso regression, also known as L1 regularisation, imposes a penalty on the magnitude of the regression coefficients. The best parameter estimate is found by solving (2.6), where $||$ denotes the absolute value. The lasso penalty often forces some of the regression coefficients to exactly zero, in effect acting as a subset selection method, discussed extensively in [28]. Whilst lasso regression is less likely to achieve better performance than ridge regression, we can examine the non-zero regression coefficients to understand which of the predictor variables are most important.

$$\hat{\boldsymbol{\beta}}^{lasso} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left[\sum_{i=1}^{n} (y^{(i)} - \beta_0 - \sum_{j=1}^{p} x_j^{(i)} \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right] \tag{2.6}$$

Chapter 3

Previous Work and Research Goals

3.1 Previous Work

As stated in the Section 2.3, there exist two broad categories for downscaling low resolution output from global atmospheric models; *statistical* and *dynamical* downscaling. In this section, we discuss previous research into combining these two methods, as this is the approach adopted in this thesis. The first step in combining these methods is application of a regional climate model (RCM) to generate high resolution reanalysis data spanning several years. Using this data, we construct a statistical relationship between the output of the large scale atmospheric model, with the RCM output.

There are two main motivations for combining these methods. By dynamically downscaling to produce several years of high resolution gridded data, we can provide statistical models with high resolution data to learn from, rather than sparse data taken from measurement stations. Secondly, we save on computational resource as we only need to produce several years of high resolution data from the RCM, in order to learn a transfer function capable of mapping low resolution reanalysis data to the resolution of the RCM. If an accurate model is constructed, we can produce high resolution gridded data, by using only low resolution data as input to our model.

The combined statistical-dynamical approach considered in this thesis was introduced recently by Huang et al.[22]. In this work they use a multivariate linear regression (MLR) model to develop a transfer function between wind reanalysis data at large (32km) and local scales (3km) with hourly frequency for a region in Southern California. Their analysis consists of first using domain specific knowledge to derive a preliminary estimate of the downscaled u, v components to account for surface rough-

ness[1] differences between the low and high resolution datasets. Using this preliminary estimate along with several other large scale meteorological variables as covariates in an unregularised model, they learn a transfer function by training to predict one year of dynamically downscaled data. Using this method they find they can accurately produce the dynamically downscaled data to within an error $< 1.5\text{ms}^{-1}$, where the largest errors are found in regions of highly complex terrain.

At the time of writing, we find no other published literature using the combined statistical-dynamical approach introduced in [22], when tasked with downscaling wind components. As mentioned previously however there have been many studies into the purely statistical approach [25, 30, 31, 27, 26]. One study of interest was conducted by Mao and Monohan [26], who have shown that to predict surface wind components at over 2000 stations across the globe, non-linear techniques such as neural networks, random forest and support vector regression offer little improvement over linear techniques such as multivariate linear regression. They do however conclude that robust methods of feature selection, such as lasso regression, can improve linear predictability when downscaling to measured data.

3.2 Research Goals

Based on previous research, we can say that a detailed exploration of methods for linear regression has not yet been applied in the statistical-dynamical approach. Whilst Huang et al. [22] have explored an MLR model and some interpretability by examining the contributions of each predictor, they do not employ feature selection methods such as lasso regression, discussed in Section 2.4.2. The lasso method allows us to interpret the most important variables in the analysis, as it will reduce the least important variable coefficients to exactly zero, in effect acting as a continuous subset selection method [28].

Moreover, there is no mention of regularisation measures to improve prediction performance, which is likely to lead to improved predictions when dealing with highly correlated inputs. Finally, not only is it of interest to build upon the work of Huang et al. [22], it is also important to test that the methodology considered in their study applies to other geographic domains. Hence, to contribute to the field by performing a detailed exploration of linear regression methods, and to understand if we can extend

[1] Surface roughness is a geometric characteristic that describes the efficiency of the surface as a momentum sink, it is often used as a measure of inhomogeneity in the land surface [29].

the high resolution data using these methods, we outline the following research goals.

- **Goal 1**: To evaluate whether multivariate linear regression outperforms a baseline bi-linear interpolation method, showing that the approach to combine statistical-dynamical downscaling introduced in [22] is appropriate for the geographic domain and data.

- **Goal 2**: To explore methods for linear regression which introduce weight penalisations to analyse the importance of the predictors (lasso regression), and to increase prediction performance (ridge regression).

- **Goal 3**: To explore the addition of auxiliary variables, such as different meteorological predictors, or dummy variables to encode the prediction month, to understand if these variables aid the models in dealing with the seasonal differences between both the low and high resolution datasets.

Chapter 4

Data

Here we provide detail on the datasets used in this study, specifically exploring the temporal and spatial variability of both the high and low resolution datasets. Further, we discuss how we subset the data into training, validation and testing sets, and discuss the specific geographic regions we consider in our analysis.

4.1 Low Resolution Wind Data

The ERA5 reanalysis data is used as the low resolution data in this study, with a spatial resolution of $\sim$(30 km $\times$ 30 km). As the motivation of this study is to extend a wind dataset such that accurate wind resource assessments can be made, we predict the wind speed at 100m which is close to hub-height for large wind turbines [32]. Hence, we use both u, v wind components at 100m from this dataset.

We see in Figure 4.1(right) that whilst this data captures the coastal boundaries relatively well, it clearly misses topological features such as mountain ranges in the north of England and Scotland. With such a low resolution, it is likely to completely miss surface inhomogeneities on the scale of up to 30 km, and therefore fail to account for local scale forces which affect near surface winds.

As well as geographical features present in the wind datasets, we explore both low and high resolution wind datasets for seasonality effects. We obtain daily statistics for the data (averaged over all gridpoints) for each year, and average these daily statistics over the training period (Figure 4.2(a)). We see clear seasonal effect in the mean and variance of u and v in the low resolution dataset, showing larger magnitudes in the colder seasons, along with larger variability in these periods.

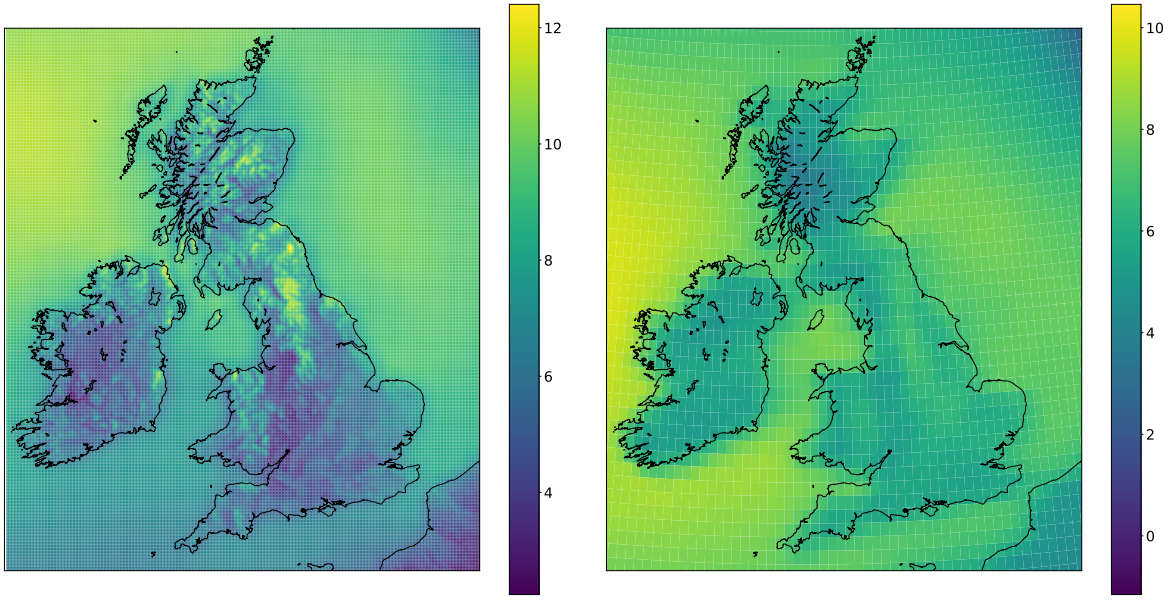


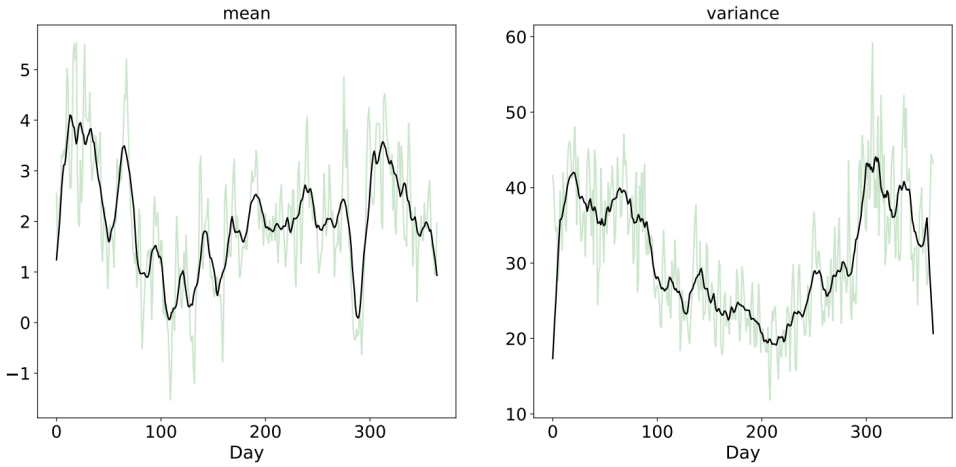
Figure 4.1: Left: High resolution data [12], right: low resolution data [10]. Both plots show the u wind component averaged over the training set period.

4.2 High Resolution Wind Data

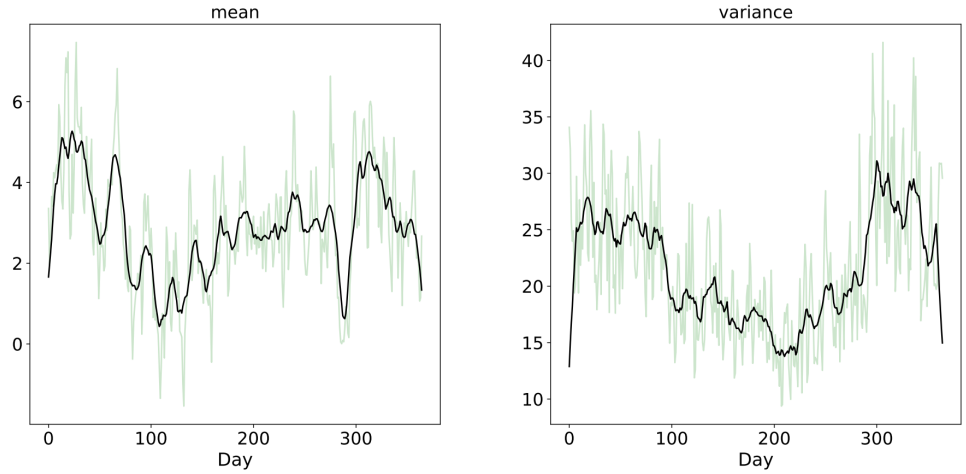
The high resolution wind data for this study comes from a regional climate model, with spatial resolution $\sim$(3 km $\times$ 3 km). As mentioned in Section 4.1, we aim to predict near hub-height at 100m and therefore restrict our analysis to the 100m data from the high resolution dataset. We see from 4.1(left), due to the finer resolution, this dataset captures some of the heterogeneity in the land surface, such as the Pennines and the mountains in northern Scotland and Wales. Much like the low resolution data we see seasonal patterns in the mean and variance of the high resolution data. Again, we see larger magnitudes in the mean during colder seasons, along with larger variability in these periods (Figure 4.2(b)). A notable difference to the low resolution data is the overall reduced variance of the high resolution dataset, with values ranging from 15-30, in comparison to a range of 20-40 for the low resolution data.

4.3 Subsetting the Data

To enable construction and evaluation of our models, we first split the data into three sets: the training data consists of a 4 year period (2001-2005), with two years for validation (2006-2008) and two years of test data (2008-2010); we omit the years 2005 and 2010 due to a small number of missing values in the dataset during these



(a) Low resolution data



(b) High resolution data

Figure 4.2: (a) Mean and variance averaged over all spatial locations and a 24 hour period for the low resolution dataset. (b) shows the same but for the high resolution data. In both figures, green shows the raw daily averaged value, and black shows a moving average of these raw values, with window size = 14, corresponding to a two week average. In all plots, day zero is January 1st.

years. Note that we only perform model training and validation on data for the u wind component. We reserve data for the v wind component as an extra test of our generalisation performance.

Further, due to time constraints, training models over the full spatial region was not feasible in the 11 week period, so we restrict the spatial domain to several separated sub-regions. To understand how the models perform in regions with highly different topography and weather patterns, we pick four regions with size $\sim (150\,\text{km} \times 150\,\text{km})$

to explore: a coastal region, a region in the Scottish Highlands, a region in the North sea and a landlocked region with relatively flat terrain (Figure 4.3). A spatial reference for these locations is shown in Appendix A.2.

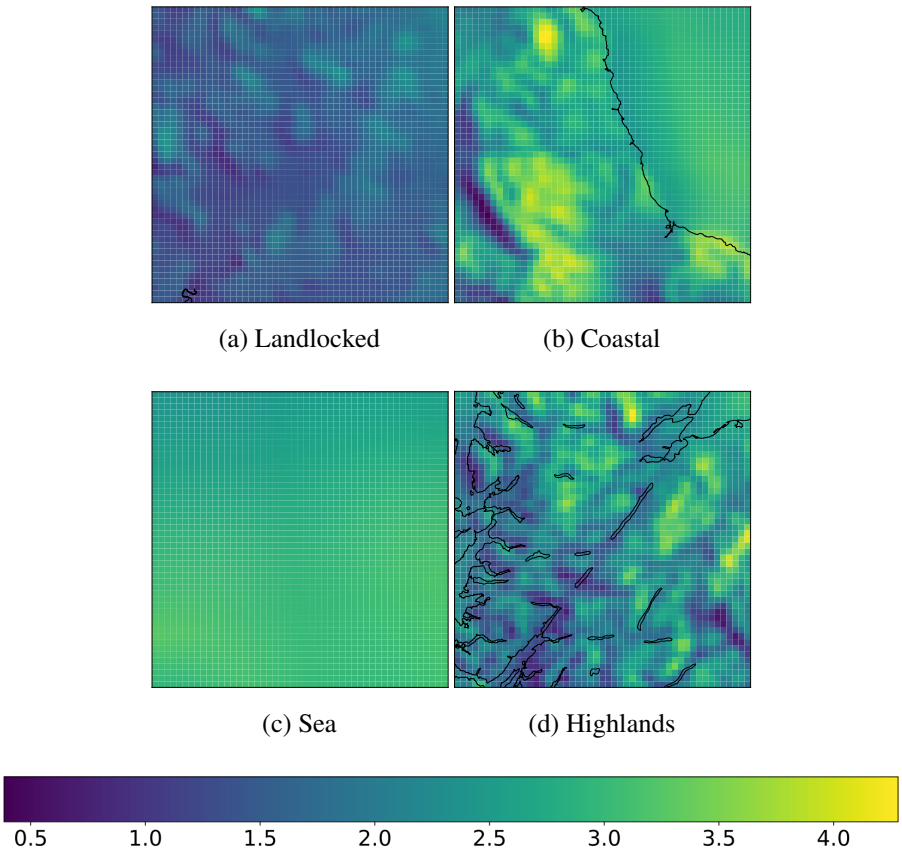


Figure 4.3: u component averaged over all hourly values in the training set for the four regions considered in the analysis.

4.4 Auxiliary Predictors

The auxiliary predictor data used in this analysis also comes from the ERA5 reanalysis dataset. As mentioned in Section 2.1, the wind is affected by many meteorological variables, so we introduce vorticity, temperature and surface pressure as auxiliary variables. We include these in the analysis based on domain specific knowledge, as these variables have direct physical influences on near surface flows [33], and for their efficacy in previous analyses [26, 16, 34, 27]. Seasonal explorations of these variables are provided in Appendix A.1.

Chapter 5

Methodology

In this chapter, we outline the methods applied to statistically downscale the low resolution wind data. As a baseline, we use simple bi-linear interpolation, often used as a quick method for downscaling [35, 36], followed by several regression methods such as multivariate linear regression (MLR), ridge regression and lasso regression. Finally, we discuss the validation metrics used to asses model performance.

5.1 Prediction Methods

5.1.1 Baseline: Bi-linear Interpolation

The most basic method often used within the literature to downscale low resolution atmospheric models is bi-linear interpolation [35, 36]. This method is a simple extension of one dimensional linear interpolation, commonly used to interpolate functions of two variables on a two dimensional grid.

In the framework of this thesis, our aim is to interpolate the low resolution u component of wind to the gridpoints of the high resolution data by evaluating some unknown function f at each of the high resolution gridpoints. We denote a point in the high resolution grid as $\ell(\phi,\theta)$ where (ϕ,θ) represent the longitude and latitude coordinates for the given point $\ell(\phi,\theta)$.

Given a point in the high resolution grid, $\ell(\phi,\theta)$, to interpolate the low resolution data to this point, we require the values of the function, f, at the locations of the four nearest neighbours of $\ell(\phi,\theta)$ in the low resolution grid. We define these as $u_{11} = f(\phi_1,\theta_1), u_{12} = f(\phi_1,\theta_2), u_{21} = f(\phi_2,\theta_1)$ and $u_{22} = f(\phi_2,\theta_2)$, with the corresponding locations (ϕ_1,θ_1), (ϕ_1,θ_2), (ϕ_2,θ_1), (ϕ_2,θ_2). To obtain the interpolated

value, we assume we know $u_{11}, u_{12}, u_{21}, u_{22}$ and must solve (5.1), where the coefficients $a_0, a_1, a_2$and a_3 are found by solving the linear system of equations in (5.2), where $f(\phi, \theta)$ denotes the desired interpolated value.

$$f(\phi, \theta) = a_0 + a_1\phi + a_2\theta + a_3\phi\theta \tag{5.1}$$

$$\begin{bmatrix} 1 & \phi_1 & \theta_1 & \phi_1\theta_1 \\ 1 & \phi_1 & \theta_2 & \phi_1\theta_2 \\ 1 & \phi_2 & \theta_1 & \phi_2\theta_1 \\ 1 & \phi_2 & \theta_2 & \phi_2\theta_2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} u_{11} \\ u_{12} \\ u_{21} \\ u_{22} \end{bmatrix} \tag{5.2}$$

Whilst the bi-linear interpolation method requires no statistical learning and is therefore quick to apply, it takes no account of the high resolution data, and only acts to smooth the values of the low resolution data to the points of the high resolution grid. Hence, it is unlikely that bi-linear interpolation will be able to pick out geographical features missed within the low resolution dataset. This motivates the need for more complex methods which incorporate information on topography, to learn topographical features which influence surface flows.

5.1.2 Multivariate Linear Regression

Another method that has been applied to this task in previous studies is multivariate linear regression, using the low resolution data as the predictor variables, and the high resolution data as the response variable. By learning a mapping from the low resolution data to predict the high resolution data, we aim to learn features that may be missed within the low resolution dataset such as surface inhomogeneity.

Suppose again we are aiming to downscale the u wind component; the regression problem is formulated as follows: Given a high resolution gridpoint $\ell(\phi, \theta)$, we wish to predict the u component value at this gridpoint at time t, denoted as y_ℓ^t. For simplicity, say we choose to use the low resolution u component as the only predictor in the model. We take N nearest neighbours to the query point, $\ell(\phi, \theta)$, from the low resolution grid at time t, which we define as $\mathbf{u}_\ell^t = [u_1(\ell, t) \quad u_2(\ell, t) \ldots \quad \ldots u_N(\ell, t)]$. Moreover, we use τ previous time points for each of the nearest neighbours giving us the following vector of covariates. To account for a bias, we insert a 1 in the first column of our covariate vector, $\mathbf{x}_\ell^t$.

$$\mathbf{x}_\ell^t = [1 \quad \mathbf{u}_\ell^t \quad \mathbf{u}_\ell^{t-1} \ldots \quad \ldots \mathbf{u}_\ell^{t-\tau}] \tag{5.3}$$

In some of our experiments, we include auxiliary variables as mentioned in Section 4.4. Using this formulation we can easily add these variables to the model. For example, to include the low resolution v component as a predictor, our covariate vector $\mathbf{x}_\ell^t$ would follow the form as in (5.4). The same applies if we wish to include the auxiliary variables such as temperature, vorticity *etc*. The full linear regression model is given in (5.5), where $\boldsymbol{\beta}$ is the vector of model parameters.

$$\mathbf{x}_\ell^t = [1 \quad \mathbf{u}_\ell^t \quad \mathbf{u}_\ell^{t-1} \dots \quad \dots \mathbf{u}_\ell^{t-\tau} \quad \mathbf{v}_\ell^t \quad \mathbf{v}_\ell^{t-1} \dots \quad \dots \mathbf{v}_\ell^{t-\tau}] \tag{5.4}$$

$$y_\ell^t = \mathbf{x}_\ell^t \boldsymbol{\beta} \tag{5.5}$$

We learn this model for each gridpoint individually, using M observations of the high resolution u component, y_ℓ^t, giving a target vector $\mathbf{y}_\ell$, and M samples of the covariates, giving the data matrix $\mathbf{X}$, both shown in (5.6).

$$\boldsymbol{y}_\ell = \begin{bmatrix} y_\ell^t \\ y_\ell^{t+1} \\ \vdots \\ y_\ell^{t+M} \end{bmatrix} \qquad \mathbf{X}_\ell = \begin{bmatrix} \mathbf{x}_\ell^t \\ \mathbf{x}_\ell^{t+1} \\ \vdots \\ \mathbf{x}_\ell^{t+M} \end{bmatrix} \tag{5.6}$$

We obtain a vector $\hat{\boldsymbol{y}}_\ell = \mathbf{X}_\ell \hat{\boldsymbol{\beta}}$ of M predictions by regressing $\mathbf{X}_\ell$ on our target vector $\mathbf{y}_\ell$ through minimisation of the least squares cost.

5.1.2.1 Lasso and Ridge Regression

Lasso and ridge regression can be applied easily within this formulation by simply adding in an additional term to the cost functions, as described in Section 2.4.2. However, to achieve optimal performance, we must optimise the penalty term, λ, for each regression model individually. In the approach considered, this would require optimisation of 2500λ's corresponding to the total number of grid points in one geographic area. This was not possible given the duration of the project, so an alternative method was used to pick the magnitude of the penalty.

We begin by randomly selecting 100 of all 2500 gridpoints for each of the four geographic regions considered (Highlands, North sea, *etc*.) For these 100 points, we perform a grid search over a log space ranging from 10×10^{-3} to 10×10^{3} to find the

optimal λ for each of the points, giving 100 optimal λ values. The optimal lambdas are chosen based on their performance on the validation set. Finally, we take the mean, μ_λ, of the exponent of the λ values, and use $\lambda = 10 \times 10^{\mu_\lambda}$ as the penalty term for the full region. Validation of this method for choosing λ is detailed in Appendix A.4.

5.1.3 Pre-processing Steps

Before applying any of the regression models, we perform feature standardization such that each column of the data matrix has zero mean and unit variance. This is of particular importance for the shrinkage methods, which are sensitive to the scale of the features.

5.2 Evaluation

In this section we discuss several metrics for evaluating the models discussed above. For a given region (coastal, North sea *etc*), our models predict hourly wind component values for each high resolution gridpoint within the given region, totalling M predictions for each gridpoint. As we are dealing with spatio-temporal data, we explore the accuracy of the models when the predictions are averaged over time, or over both space and time.

5.2.1 Temporal Errors

To explore if the models are affected by seasonality, we find the mean daily absolute error, where the mean is taken over all the gridpoints in the given region, over a 24 hour period, and over both validation years. We evaluate the error defined by (5.7) where L is the number of points within the considered region, $||$ denotes the element-wise absolute value operator, and $\boldsymbol{y}_\ell^{t,day}$, $\hat{\boldsymbol{y}}_\ell^{t,day}$ denote the ground truth and predictions of a given day in both validation years.

$$\text{MAE}^{day} = \frac{1}{\ell}\sum_{\ell}^{\ell}\frac{1}{48}\sum_{t}^{48}\left|\boldsymbol{y}_\ell^{t,day} - \hat{\boldsymbol{y}}_\ell^{t,day}\right| \tag{5.7}$$

5.2.2 Total Error

To obtain a succinct description of model performance for a given region, we calculate the total relative RMSE by averaging the relative RMSE over space. This error is a

single number describing the total relative error of the model in the given region, over the prediction period. The reason we take relative errors is because the magnitude of the wind components significantly varies from region to region, meaning we must discuss relative errors if we are to compare models trained at different locations. It is common to use the mean as a normaliser, however since the wind components u,v are directional, they can take positive or negative values, often resulting in a near zero mean. Hence, we use the interquartile range (IQR), as it is a robust measure of scale that is insensitive to the sign of the data. To obtain the normaliser, the we take the IQR of the ground truth values over all the gridpoints in the region and the M hours predicted. The full error metric is given by (5.8), where L denotes the number of gridpoints in the region, M denotes the number of hours predicted and y_ℓ^t,$\hat{y}_\ell^t$ denote the ground truth and predicted value at time t at gridpoint l.

$$\text{Total Relative RMSE} = \frac{\sqrt{\frac{1}{M}\frac{1}{\ell}\sum_\ell^L \sum_t^M (y_\ell^t - \hat{y}_\ell^t)^2}}{IQR} \tag{5.8}$$

5.2.3 Evaluating the Generalisation Performance

Due to the large volume of data at our disposal, we explore two methods to evaluate the generalisation performance of our models. As is common amongst almost all machine learning studies, we save a test set for the variable we are predicting, namely the u wind component, to assess whether the selected models obtain high performance on unseen data.

Further, we have many years of data for the v wind component and can therefore also asses our model choices based on our prediction performance on the v wind component. We note that in order to do this, we retrain the models with the v component as the target data [1], however we do not re-optimise the regularisation constant λ, and use the same λ found when training to predict the u component. Note that here we are not testing the generalisation performance of the model, we are understanding whether the model choices made during investigation on the u component, also allow us to achieve low errors when predicting the v component.

[1] We use the same training-test set split; the training set covers 2001- 2005, and the testing set covers 2008-2010

Chapter 6

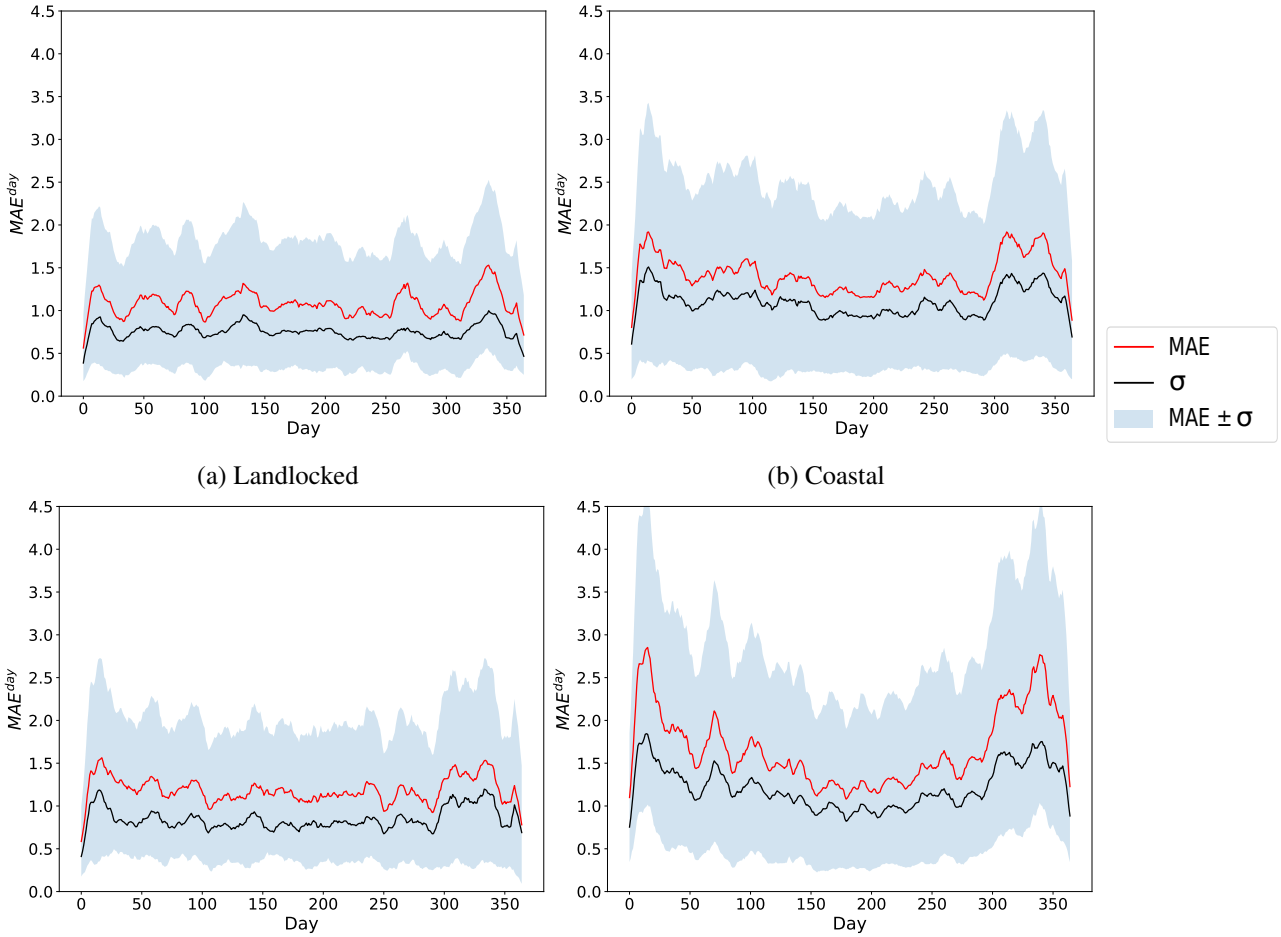
Results

In this chapter we show the results of the experiments conducted to achieve the goals outlined in Section 3.2. Firstly we validate the approach proposed in [22], showing that the proposed method is valid for this dataset, and outperforms the baseline bi-linear interpolation. Secondly, we perform an in-depth study into the applications of several regression methods, allowing us to obtain improved prediction performance, and interpret the most useful variables present in the models. We conclude this chapter with evaluation of the models on the testing sets.

6.1 Baseline: Bi-linear Interpolation

Table 6.1 shows the total relative RMSE for the u wind component using the bi-linear interpolation method. We see that regions with complex terrain are more difficult to predict, with the Highlands region having 13% higher error than the error over the sea. The bi-linear interpolation method is in effect a distance weighted average of the four points used for interpolation, and hence acts as to smooth the data. This is clear from Figure 6.2 and is to be expected, as the low resolution data is simply to coarse to account for topographic features on the scale of hills and valleys.

We find a clear seasonal effect in the daily averaged mean absolute error (MAE^{day}) for the Highlands region and observe consistently larger mean and standard deviations of the daily error values during the colder periods when compared to the warmer months (Figure 6.1). This is most likely because of the lower variance in the wind components during the warmer periods, present in both datasets.



(a) Landlocked

(b) Coastal

(d) Sea

(e) Highlands

Figure 6.1: Mean absolute error averaged over all gridpoints and a daily period in both years of the validations set, (MAE^{day}), using the bi-linear interpolation method. σ denotes the standard deviation over this period. Note the clear seasonal effect in the MAE, and σ for the highlands region.

Region	Landlocked	Coastal	Sea	Highlands
Interpolated	0.2159	0.2210	0.1718	0.3019
MLR	**0.1757**	**0.1781**	**0.1547**	**0.2219**

Table 6.1: Total relative RMSE on the validation set for the u wind components using the baseline bi-linear interpolation method and the simple multivariate linear regression.

6.2 Multivariate Linear Regression

To understand the effect of more complex models, we first experiment with a multivariate linear regression (MLR) model using the same data as used in the interpolation

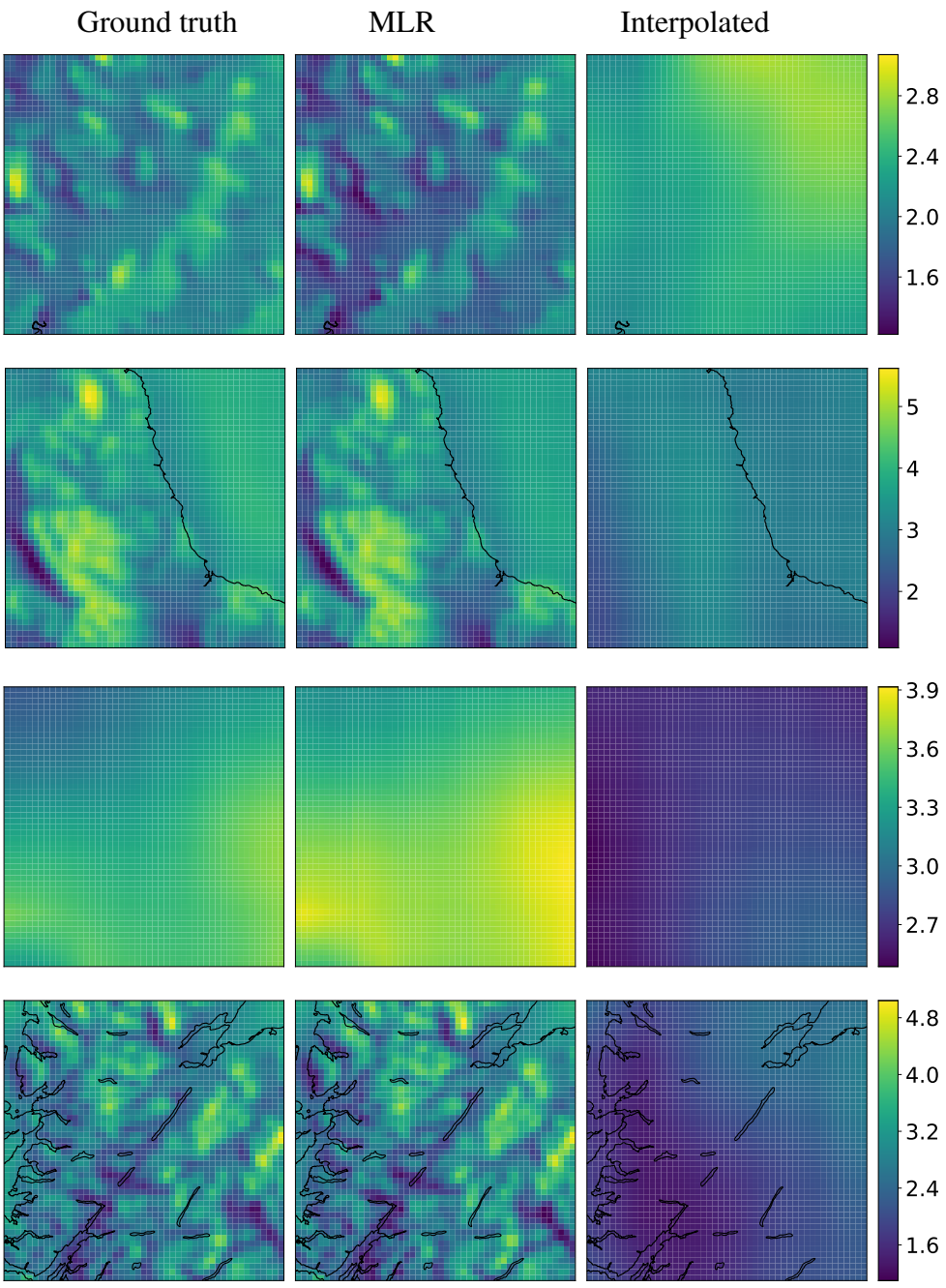
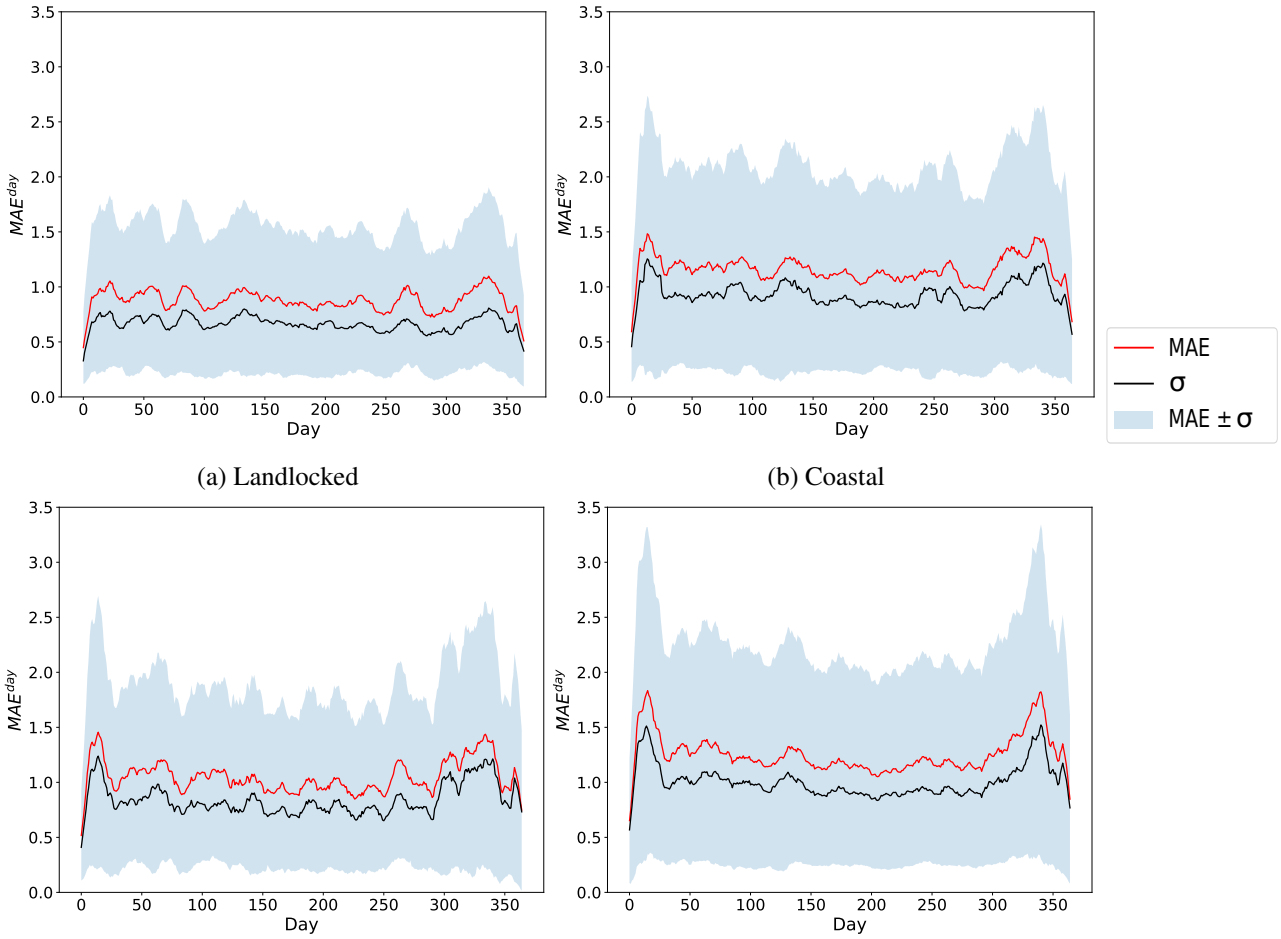


Figure 6.2: Left column shows the ground truth high resolution data, middle column shows the results of the simple multivariate linear regression model, and right shows the results of the bi-linear interpolation. All plots show the u component, averaged over the validation set period. Notice how the bi-linear interpolation acts to smooth the data, whilst the MLR model is able to model some of the surface inhomogeneity in the ground truth data.

method. This corresponds to using 4 nearest neighbours, i.e $N = 4$ and no previous time points ($\tau = 0$). We use no weight penalisation for this experiment. We see from Table 6.1 that even this simple MLR model offers great improvements over the baseline method, reducing the error from $\sim 30\%$ to $\sim 22\%$ (in the Highlands). Through the inclusion of the high resolution data as ground truth, it is clear that the model is able to learn some of the geographic features present within the data (Figure 6.2). Further, the seasonal structure within the MAE^{day} appears somewhat reduced, compared to the bi-linear interpolation method (Figure 6.3).



(a) Landlocked

(b) Coastal

(d) Sea

(e) Highlands

Figure 6.3: Shows the mean absolute error averaged over all gridpoints and a daily period in both years of the validations set, (MAE^{day}), using the simple MLR model described in Section 6.2. σ denotes the standard deviation over this period.

6.3 Ridge Regression

The next set of experiments explore the effects of additional covariates using ridge regression to prevent overfitting due to the addition of many more features. We optimise the regularisation coefficient λ using the method discussed in Section 5.1.2.1. First, we explore the effects of increasing N, the number of nearest neighbours to our query point, whilst holding the number of previous time points at $\tau = 0$. Then, we experiment with increasing τ whilst holding $N = 4$. All experiments in this section use no auxiliary variables, and the only covariates are the nearest neighbours or previous time points of the u component.

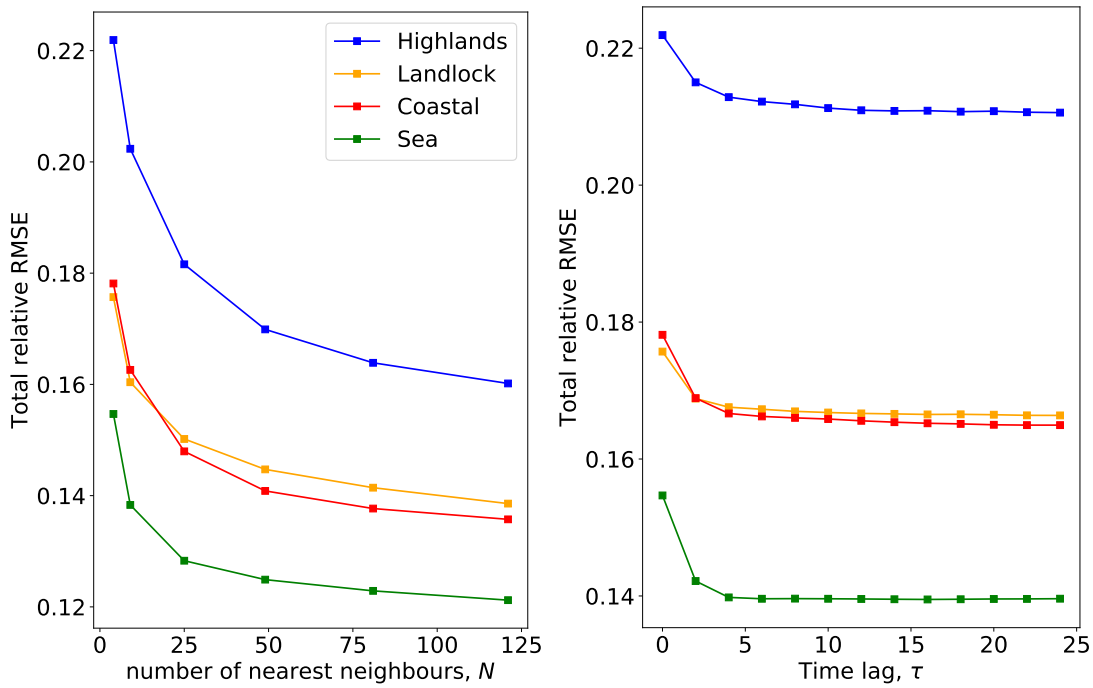
6.3.1 Increasing the Number of Nearest Neighbours N

By increasing the neighbourhood of nearest neighbours, we expect that prediction performance will improve as we provide the model with a more holistic picture of the state of the wind climate at a given time. As the low resolution components are located on a two-dimensional grid and the distance we use is the intrinsic distance[1], there are 9 neighbours at distance one; 25 neighbours at distance two; 49 neighbours at distance three, and so on. Thus, we run experiments with $N = 9$ to $N = 121$ (Figure 6.4)(a). We find that when using $N = 49$ we lower the total relative RMSE by ($\sim 5\%$) over the results when using $N = 4$. However, increasing N to 121, yields $< 1.0\%$ improvement at a high computational cost from the addition of 72 extra features (see Appendix A.3).

6.3.2 Increasing the Number of Previous Time Points τ

We now explore the effect of increasing τ, the number of historic time points used as features in the model. We expect that increasing τ will increase prediction performance to some extent as this provides information on the temporal variability of the component, prior to the prediction hour. As mentioned above, we hold $N = 4$ for all experiments, use ridge regression, and incrementally increase τ in steps of 2, from 2 up to 24. Figure 6.4(b) shows that using 5 previous time points reduces the error by 1% after which the performance increase begins to plateau. This may indicate that correlations between the low and high resolution wind components are dampened after a 5 hour period, implying that using historic data beyond a 5 hour period will not be

[1] Intrinsic distance in this context refers to the distance in the grid, i.e the distance that is intrinsic to the space which the data occupy.



(a) (b)

Figure 6.4: (a) Shows the total relative RMSE on the validation set as we change the number of nearest neighbours used as covariates in the model, whilst holding $\tau = 0$. (b) Shows the same as (a) however, N is now set to 4, and we vary τ from 2-24.

useful in further analysis and hence we restrict $\tau = 5$ for further experiments.

6.4 Adding Auxiliary Variables with Optimal N and τ

The next experiment aims to show the effects of adding in auxiliary variables which may contain predictive information on forces which influence both wind components. As a comparison, we compare against the results obtained when using only u as the predictor variable. In all experiments, we use $N = 49$ and $\tau = 5$, based on the results of the previous experiments, and re-optimise λ for each separate experiment. Further, we restrict our analysis to the most complex region, i.e the Highlands, as the previous experiments show this to be the most difficult region to predict. This is evident from the higher total relative RMSE values, and a clearly seasonal pattern present in the mean absolute errors.

In addition to adding in auxiliary meteorological variables, we experiment with the inclusion of dummy variables to encode the month of the year. This adds an additional

11 variables, where all are zero except one, indicating the month for each sample point. The intention is to target the seasonality effect within the errors, to give the model some temporal information which may permit learning the differences between the seasonal patterns observed within the low and high resolution data.

We see that when using optimal N and τ, the seasonality effect is again reduced as the difference in the MAE $+\sigma$ between the cooler and warmer period has dropped from ~ 1.5 (Figure 6.3) to ~ 0.5 (Figure 6.5). Adding in the auxiliary meteorological variables reduces the total relative RMSE by 1.65% (Table 6.2), yet still exhibits a small seasonality effect, with spikes in the mean MAE and σ in the cooler periods. Surprisingly, inclusion of the dummy variables to encode the month of year has had a negative impact on the performance (Table 6.2), and has not reduced the seasonality effect. It is possible that a monthly encoding is too complex, as the pattern we observe operates over a seasonal scale rather a monthly time scale. Further analysis could experiment with encoding the prediction season, instead of prediction month.

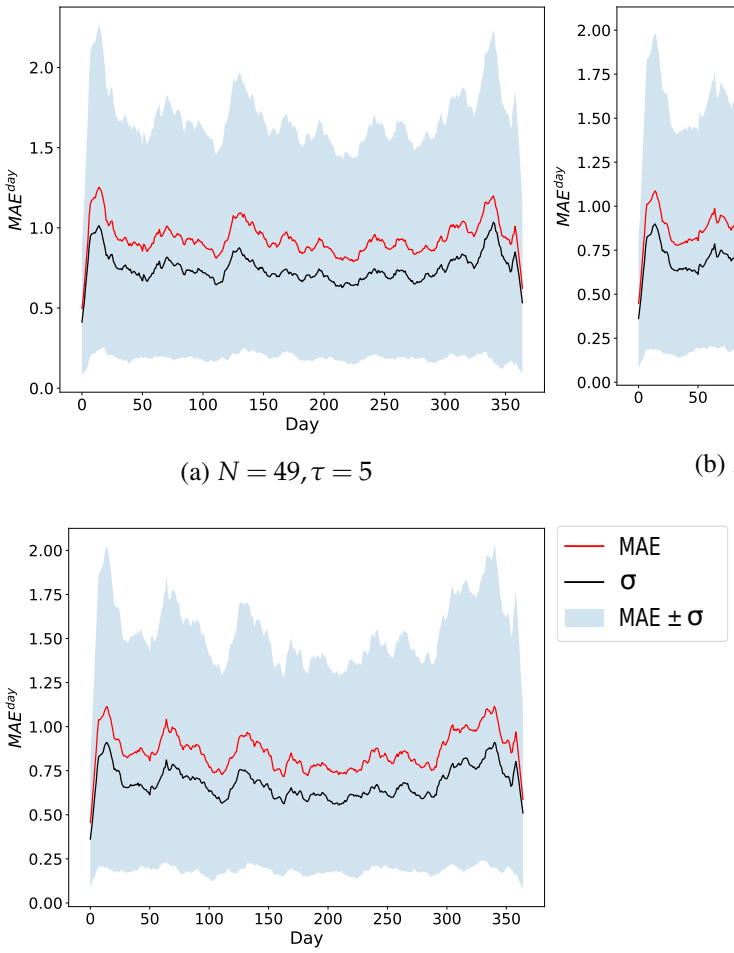
Model	$N = 49, \tau = 5$	$N = 49, \tau = 5 +$ aux	$N = 49, \tau = 5 +$ aux + dummy variables
TRRMSE	0.1621	**0.1456**	0.1491

Table 6.2: Shows the total relative RMSE (TTRMSE) on the validation set for the MLR models with three sets of covariates, given in row 1. + aux implies inclusion of auxilliary variables, and +dummy variables implies inclusion of one-hot encoded variables to indicate the prediction month.

6.5 Identifying the Important Variables

Having determined which variables offer substantial performance increase, we experiment with L1 regularisation (lasso regression) to perform feature selection. This will allow us to understand which variables are the best predictors of the u wind component, and experiment with models that exclude the least important predictors, thus reducing the computational cost of prediction. From the results of previous experiments, we have found the optimal number of nearest neighbours, $N = 49$, and the optimal number of previous time points, $\tau = 5$. Further, the result of experiment 6.4 shows that using the auxiliary variables does increase prediction performance considerably. However, using all of these variables totals $((5 + 1) * 49) * 5 = 1470$ covariates in our models [2].

[2]This number is calculated using the formula:
$((\tau + 1) * N) *$ number of different meteorological variables)



(a) $N = 49, \tau = 5$

(b) $N = 49, \tau = 5$ + aux

(c) $N = 49, \tau = 5$ + aux + dummy variables

Figure 6.5: Shows the mean absolute error averaged over all gridpoints and a daily period in both years of the validations set, (MAE^{day}). σ denotes the standard deviation over this period. (a) errors using MLR model with optimal N and τ, (b) is (a) with the inclusion of the auxiliary variables and (c) is (b) with the inclusion of one hot encoded variables for the month of the year.

As our features are closely connected in time and space, it is likely that there are large correlations between many of the features in our model. We can determine the prevalence of these correlations by performing principal component analysis (PCA) on the data matrices used in the regression models. We find that to retain 99% of the variance, we only require 49 principal components (see Appendix A.5), confirming that our data are highly correlated, and it is possible that retaining a small subset of our features will achieve high performance.

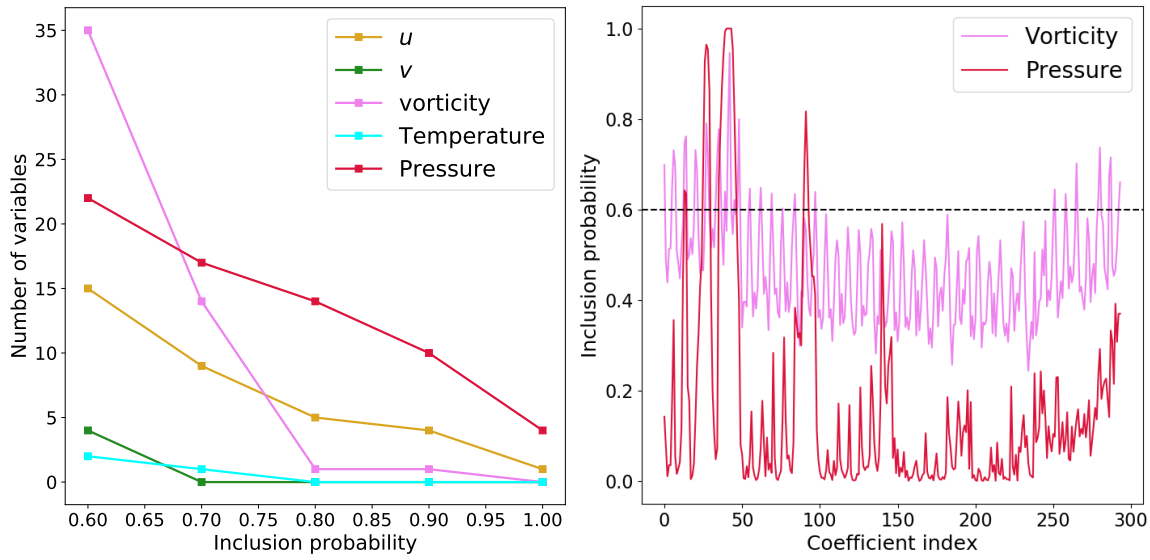
6.5.1 Finding the Most Important Meteorological Variable

To determine the most important meteorological variables we perform an experiment using lasso regression, with τ and N held at the optimal values, and use all five meteorological variables. We optimise lambda using the method described in Section 5.1.2.1 and record the learned weights of all 2500 regression models we construct for the Highlands region.

Using all weight vectors for all 2500 regression models, we compute the *inclusion probability* of a variable. This is defined as the total number of times the coefficient associated with a given feature is non-zero, normalised by the total number of regression models. To determine the importance of the different types of meteorological variable, we explore groups of variables grouped by meteorological type (u,v,vorticity *etc.*). Note that all groups have equal size and therefore we can fairly compare the variable importance based on the number of included variables from each group.

Figure 6.6(a) shows that the u wind component, and pressure are selected most often, whilst temperature and the v wind component are selected the least. We see that in 20% of the models, no temperature or v variables are selected. Further, we observe a large decrease in the number of included vorticity variables as we reach 80% inclusion probability. This is likely due to the a high mean inclusion probability for the vorticity variables (Figure 6.6(b)) indicating that there is no clear preference for a specific vorticity variable, and any of them have $\sim 50\%$ probability of inclusion. This is in contrast to the pressure variables, where there is clear preference, indicated by large spikes in the inclusion probability for a specific set of the pressure variables. (Figure 6.6(b)).

To identify the most important meteorological variable in predicting the u component, we look at those features present within 100% of the models, thus having an inclusion probability equal to one. Of the six variables with this probability, five are pressure variables and only one variable corresponds to the u component. This indicates that to predict the u component, pressure variables are better predictors than the u component itself. Whilst this is somewhat surprising, pressure systems are a driving force of surface flows and pressure fields have been shown in previous studies to be useful predictors of wind components [16]. From this experiment we can conclude that to predict the u wind component, data for the pressure and the component itself are most important, whilst data for temperature and the v component offer little predictive information.



(a) (b)

Figure 6.6: (a) Shows the number of included variables against increasing inclusion probability. As an example, with probability 0.6, 15 u variables are selected, implying that in 40% of the models, no more than 15 u variables have non-zero regression coefficients. (b) Shows the inclusion probability for all variables in the pressure (red) and vorticity groups (pink).

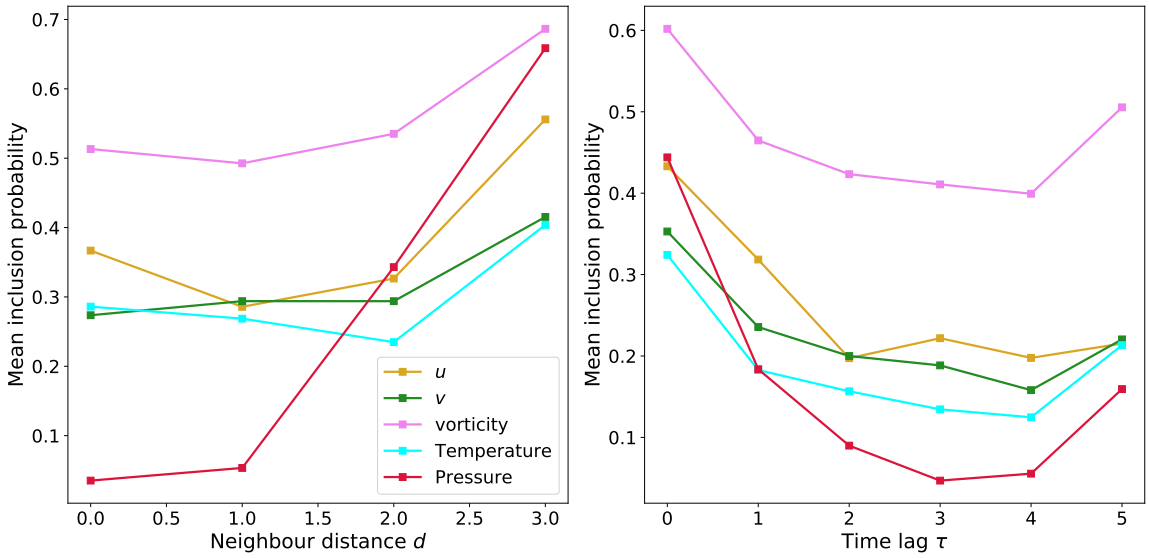
6.5.2 Identifying the Best Prediction Distance and Time

Not only is it of interest to understand which of the meteorological variables are most important, it is also instructive to understand which previous time points, indexed by τ, and neighbour distance, d, are most often included, and therefore most useful in the analysis. We group the variables by meteorological type and nearest neighbour distance, d, and take the mean inclusion probability of all variables in the group. Note that as there are many more variables at distance $d = 3$, than at $d = 0$, we must take the mean of each group to enable fair comparison between the groups.

We see that neighbours at distance $d = 3$ have the highest mean inclusion probability for all meteorological variable types (Figure 6.7(a)). Whilst one may expect that predictors close to our prediction location should be selected most often, it has been observed in other studies that this may not be the case. Curry et al. [33] have shown that observed variance in wind speeds often arises from predictor behaviour that is not necessarily the closest predictor to the point of interest. They find that in all cases, higher R^2 values are found when using predictors that are not located at the nearest

possible gridbox, which is in agreement with the results of our experiment.

Moreover, when grouping the variables by meteorological type and time lag τ, again taking the mean over the groups, we see that the variables picked most often correspond to $\tau = 0$, and interestingly, $\tau = 5$. This is likely because the feature selection process chooses to keep those variables which are least correlated, which would correspond to those separated by the largest time period. This informs us that in future study, it may be wise to experiment with predictors with a lower than hourly frequency, to reduce the correlations between inlcuded variables.



(a) (b)

Figure 6.7: (a) Shows the mean inclusion probability for the variables grouped by meteorological type and neighbour distance. (b) Shows the same however the grouping is now by meteorological type and time lag τ. The same color shceme applies to both figures and is shown in (a).

6.6 Experimenting with the Most Important Variables

Results of the previous experiments, and principal component analysis of our data, show that it is likely that we have many redundant features due to high spatial and temporal correlations between them. Here, we perform experiments using a subset of the features to understand how model performance varies as we increase the number of features in our model. We select the most important variables as those with highest

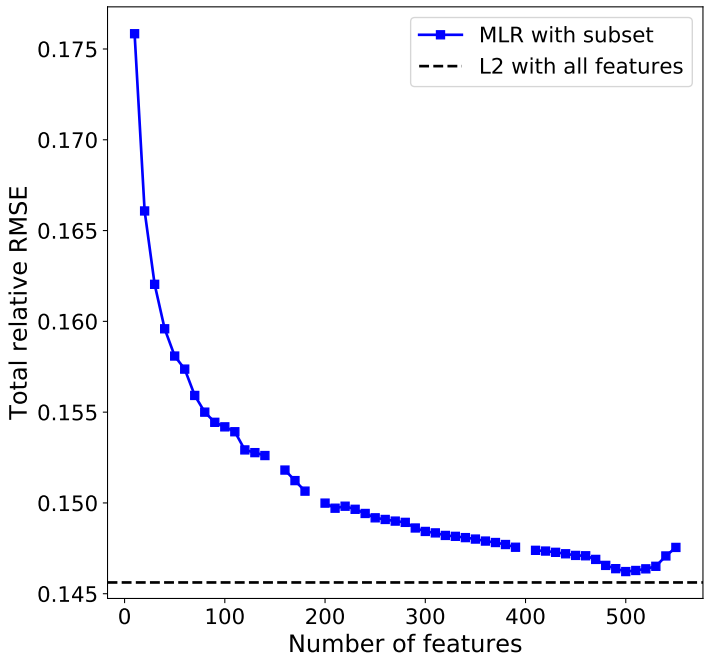


Figure 6.8: Blue shows the Total relative RMSE on the validation set as we vary the number of features included in the MLR model, the black dotted line shows the optimal performance obtained using ridge regression and all 1470 features.

inclusion probability and include these in increments of 10 variables at a time from 10-500 features. We use a multivariate linear regression model with no weight penalisation for the following experiment.

We find that to reach to within 0.005 of the optimal error value obtained using all features, we need only use 200 features (Figure 6.8). Further we see that we can obtain very close to optimal performance using 500 of the most important features, before we begin to overfit to the to the training data. From these results we see that using all 1470 covariates in an MLR model is likely too complex for the problem, adding additional computational complexity for very marginal gain.

6.7 Testing the Final Models

For our final results, we test the generalisation performance of a selection of the models by presenting their performance on both testing sets. We show a selection of the models ranging from the simple baseline bi-linear interpolation, to the most complex, namely ridge regression with all 1470 features. We see that on both testing sets, the observed behaviour of the models is in general very similar (Figure 6.9). The most

complex model, using ridge regression and all 1470 features achieves the lowest total relative RMSE, whilst an unregularised MLR model with only 500 of the most important features reaches close to optimal performance, in both cases to within $\sim 1\%$. If performance is our primary concern it is clear that we should use ridge regression with all possible features. However, the extra computational complexity associated with an additional 970 features may not be worth the $\sim 1\%$ improvement in total relative RMSE.

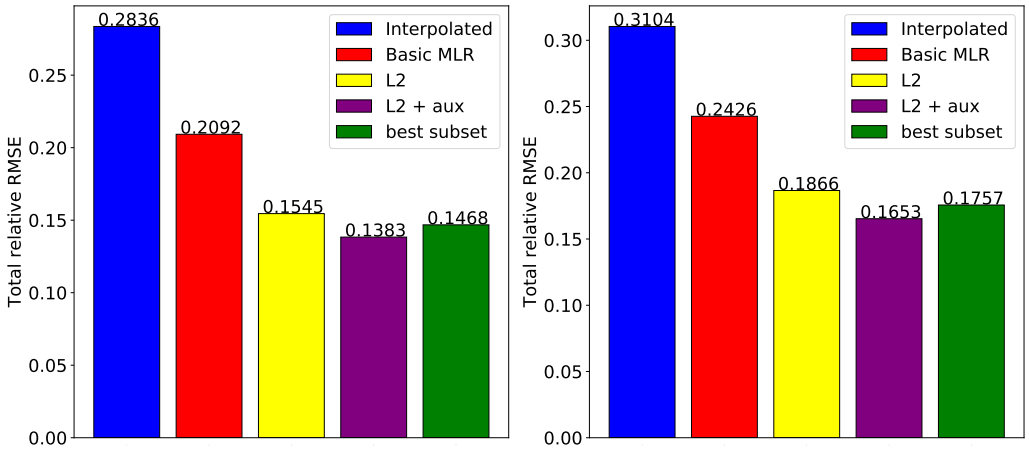


Figure 6.9: Shows the results of a selection of the models on the testing set for both u(a) and v(b) components. Blue and red show the interpolation and simple MLR model. Yellow shows the ridge regression result using only u or only v as a predictor. Purple shows the ridge regression result with all 1470 variables, and green indicates an unregularised MLR model using the 500 most important variables found through lasso regression.

Whilst we find the most complex model offers the best performance on both testing sets, it is interesting that the model selection process has generalised to the v component, considering all model choices were made solely on the u component. Further, we see that the best subset found for u also performs well when trying to predict v, indicating that the auxiliary variables useful for predicting u, are also useful for predicting v. Whilst we cannot conclude that the models we have applied to the v component are the best possible models, as we have not tuned any of the hyper-parameters to this data, we can conclude that in order to achieve reasonably good performance, offering $\sim 15\%$ improvement over the baseline method, it is possible to make model choices based only on the u component.

Chapter 7

Concluding Remarks

7.1 Summary and Conclusions

To obtain accurate assessments of United Kingdom wind resource, we require a high resolution wind dataset of sufficient temporal extent. As discussed in this work, there are high computational costs associated with dynamical downscaling, and downsides to purely statistical modelling due to scarcity of data. Combination of these methods allows us to limit the time spent on dynamical downscaling, and does not require measured data to learn a transfer function capable of mapping low resolution to high resolution data. Hence, this thesis has explored methods which combine the dynamical and statistical approach. We have conducted a detailed exploration of several methods of linear regression for the task of combined statistical-dynamical downscaling of wind components for the British Isles and surrounding water. We have shown that the combined approach to downscaling, introduced by Huang et al. [22], greatly improves over a commonly used method for increasing spatial resolution, namely bi-linear interpolation, and completes our first research goal.

The second research goal; to implement weight penalisation methods to analyse predictor importance and improve prediction accuracy, was addressed through firstly implementing ridge regression with no auxiliary variables where we find that this regression method allows us to reduce the total relative RMSE from ($\sim 22\%$ to 16%), when compared to a simple unregularised regression model. To analyse predictor importance we implemented lasso regression as a feature selection method. We find clear preference for pressure variables as predictors of the u wind component over temperature, vorticity and the v wind component. We build upon the second research goal, by performing subset analysis using features based on their importance, which was

defined by the inclusion probability. We find that using only 500 features allows us to achieve within 1% of the best performance when using all 1470 features.

We address the final research goal; to explore the addition of auxiliary variables in the L2 regularised regression model to capture the seasonality differences between the high and low resolution datasets. Whilst we find that inclusion of the auxiliary meteorological variables improves prediction performance, we still observe some seasonal structure with the mean absolute errors. Addition of variables to encode the prediction month, did not improve performance, and more study is required to understand how best to deal with seasonal differences between the high and low resolution datasets considered in this study.

7.2 Limitations and Future work

Whilst we have accomplished the goals set out in Section 3.2, there are several limitations and avenues of exploration not covered within this thesis. Firstly, a limiting factor of this study is the restriction of in-depth analysis to only one region, due to time constraints. Implementation of lasso regression to other geographical areas may reveal that the most important predictors depends on the geographic region we consider. As the atmospheric dynamics greatly differs between the four locations considered in this study, it is possible that temperature or vorticity fields, coupling near surface and upper atmospheric flows, may be better predictors for the u wind component in different locations.

A further limitation of this work, is that we have not validated the predictions against observed wind data. The models constructed in this work were tasked with predicting high resolution reanalysis data, and we have achieved this to high accuracy. The ultimate goal however, is to understand how our models perform on predicting observed wind data, such that they accurately represent the wind climate. Understanding how our models perform at producing observed wind data, would set the performance of our models within a wider context. This would allow us to assess the potential for accurate extension of the high resolution dataset, such that we are not only accurate at predicting the model, but also accurate to predicting measured wind data.

Whilst this is a limitation of our study, we note that the errors observed when predicting the high resolution data may not necessarily be a limitation of the applied regression methods. These methods may prove to achieve high performance when validating on observed wind data. However, to provide such an evaluation, we require

measured data at the prediction height of 100m. Due to lack of available data we were not able to perform this evaluation.

With regards to future study, a method to addressing the high correlations between the features is to use PCA regression. This method consists of performing regression using k principal components of the data, where k is often chosen to be the number of components that retains a high percentage of the variance within the data. As we have found through PCA on our data (see Appendix A.5), we require only 49 principal components to retain 99% of the variance of our data. Hence, performing PCA regression with $k = 49$, may offer further computational speed up compared to the ridge and lasso regression methods, without incurring a large increase in prediction error. Preliminary analysis of this method was conducted, indicating comparable results could be obtained with this method, however a full rigorous analysis was not performed due to time constraints.

Moreover, it would be instructive to explore methods such as multiple output regression or parameter sharing between adjacent grid points, to reduce the computational complexity of our methods. As the covariate matrix, $\boldsymbol{X}$, consists of nearest neighbours within the low resolution dataset, we observed that many adjacent gridpoints share the same covariate matrix, with only the target value changing between gridpoints. Multiple output regression capitalises on shared covariate matrices, performing expensive inversion of the gram matrix, $\boldsymbol{G} = \boldsymbol{X}^\top \boldsymbol{X}$, only once, over the regions where the covariate matrix is unchanged. Parameter sharing acts to take an average across the gridpoints of which we share parameters and therefore may be particularly applicable in open water regions, where there are small changes in surface topography over large spatial scales. Due to the high spatial correlation of the u component over the 3 km distance between high resolution gridpoints, it is likely that these methods may reduce the computational effort required to generate predictions for all gridpoints within the given regions, without having substantial effect on the overall performance.

Bibliography

[1] Gareth P Harrison, Samuel L Hawkins, Dan Eager, and Lucy C Cradden. Capacity value of offshore wind in Great Britain. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, 229(5):360–372, 2015.

[2] Anuj Banshwar, Naveen Kumar Sharma, Yog Raj Sood, and Rajnish Shrivastava. Renewable energy sources as a new participant in ancillary service markets. *Energy Strategy Reviews*, 18:106–120, 2017.

[3] George Crabtree et al. Integrating renewable electricity on the grid. In *AIP Conference Proceedings*, volume 1401, pages 387–405. AIP, 2011.

[4] Bastien Alonzo, Riwal Plougonven, Mathilde Mougeot, Aurélie Fischer, Aurore Dupré, and Philippe Drobinski. From Numerical Weather Prediction Outputs to Accurate Local Surface Wind Speed: Statistical Modelling and Forecasts. In *Forecasting and Risk Management for Renewable Energy*, pages 23–44. Springer, 2017.

[5] Department of Energy and Climate Change. The UK Low Carbon Transition Plan, 2009.

[6] WMO. WMO Guidelines on the Calculation of Climate Normals. 2017. ISBN 978-92-63-11203-3.

[7] Anthony Arguez, Imke Durre, Scott Applequist, Russell S Vose, Michael F Squires, Xungang Yin, Richard R Heim Jr, and Timothy W Owen. NOAA's 1981–2010 US climate normals: an overview. *Bulletin of the American Meteorological Society*, 93(11):1687–1697, 2012.

[8] Derek van der Kamp, Charles L Curry, and Adam H Monahan. Statistical downscaling of historical monthly mean winds over a coastal region of complex terrain. ii. predicting wind components. *Climate dynamics*, 38(7-8):1301–1311, 2012.

[9] Met Office Surface Data Users Guide. `https://artefacts.ceda.ac.uk/badc_datadocs/ukmo-midas/ukmo_guide.html#3.1`, 2019. Accessed: August 2019.

[10] Copernicus Climate Change Service (C3S). ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate, 2017. Accessed 2019-06 - 2019-08.

[11] Ronald Gelaro, Will McCarty, Max J Suárez, Ricardo Todling, Andrea Molod, Lawrence Takacs, Cynthia A Randles, Anton Darmenov, Michael G Bosilovich, Rolf Reichle, et al. The modern-era retrospective analysis for research and applications, version 2 (MERRA-2). *Journal of Climate*, 30(14):5419–5454, 2017.

[12] Samuel Lennon Hawkins. *High resolution reanalysis of wind speeds over the British Isles for wind energy integration*. PhD thesis, The University of Edinburgh, 2012.

[13] Warren M Washington, Lawrence Buja, and Anthony Craig. The computational future for climate and earth system models: on the path to petaflop and beyond. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1890):833–846, 2008.

[14] European Centre for Medium-Range Weather Forecasts. Annual report 2015, 2016.

[15] John M Wallace and Peter V Hobbs. *Atmospheric science: an introductory survey*, volume 92. Elsevier, 2006.

[16] SC Pryor, Justin T Schoof, and RJ Barthelmie. Empirical downscaling of wind speed probability distributions. *Journal of Geophysical Research: Atmospheres*, 110(D19), 2005.

[17] Edward N Lorenz. Reflections on the conception, birth, and childhood of numerical weather prediction. *Annu. Rev. Earth Planet. Sci.*, 34:37–45, 2006.

[18] Jon Olauson. ERA5: The new champion of wind power modelling? *Renewable energy*, 126:322–331, 2018.

[19] Edinburgh Parallel Computing Centre. `https://www.epcc.ed.ac.uk/projects-portfolio/hector-national-supercomputing-service`. Accessed July 2019.

[20] U Heikkilä, A Sandvik, and Asgeir Sorteberg. Dynamical downscaling of ERA-40 in complex terrain using the WRF regional climate model. *Climate dynamics*, 37(7-8):1551–1564, 2011.

[21] Kristian Horvath, Darko Koracin, Ramesh Vellore, Jinhua Jiang, and Radian Belu. Sub-kilometer dynamical downscaling of near-surface winds in complex terrain using wrf and mm5 mesoscale models. *Journal of Geophysical Research: Atmospheres*, 117(D11), 2012.

[22] Hsin-Yuan Huang, Scott B Capps, Shao-Ching Huang, and Alex Hall. Downscaling near-surface wind over complex terrain using a physically-based statistical modeling approach. *Climate dynamics*, 44(1-2):529–542, 2015.

[23] Joseph M Russo and John W Zack. Downscaling GCM output with a mesoscale model. *Journal of Environmental Management*, 49(1):19–29, 1997.

[24] Adam H Monahan. Can we see the wind? statistical downscaling of historical sea surface winds in the subarctic northeast pacific. *Journal of Climate*, 25(5):1511–1528, 2012.

[25] Adam Winstral, Tobias Jonas, and Nora Helbig. Statistical downscaling of gridded wind speed data using local topography. *Journal of Hydrometeorology*, 18(2):335–348, 2017.

[26] Yiwen Mao and Adam Monahan. Linear and nonlinear regression prediction of surface wind components. *Climate dynamics*, 51(9-10):3291–3309, 2018.

[27] T Salameh, P Drobinski, M Vrac, and P Naveau. Statistical downscaling of near-surface wind over complex terrain in southern france. *Meteorology and Atmospheric Physics*, 103(1-4):253–265, 2009.

[28] Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.

[29] Roland B Stull. *An introduction to boundary layer meteorology*, volume 13. Springer Science & Business Media, 2012.

[30] Megan C Kirchmeier, David J Lorenz, and Daniel J Vimont. Statistical downscaling of daily wind speed variations. *Journal of Applied Meteorology and Climatology*, 53(3):660–675, 2014.

[31] Annemarie Devis, Nicole PM van Lipzig, and Matthias Demuzere. A new statistical approach to downscale wind speed distributions at a site in northern europe. *Journal of Geophysical Research: Atmospheres*, 118(5):2272–2283, 2013.

[32] Ryan Wiser, Mark Bolinger, et al. 2018 Wind Technologies Market Report. Technical report, US Department of Energy: Office of Energy Efficiency & Renewable Energy, 2018.

[33] Charles L Curry, Derek van der Kamp, and Adam H Monahan. Statistical downscaling of historical monthly mean winds over a coastal region of complex terrain. i. predicting wind speed. *Climate dynamics*, 38(7-8):1281–1299, 2012.

[34] Aaron MR Culver and Adam H Monahan. The statistical predictability of surface winds over western and central canada. *Journal of Climate*, 26(21):8305–8322, 2013.

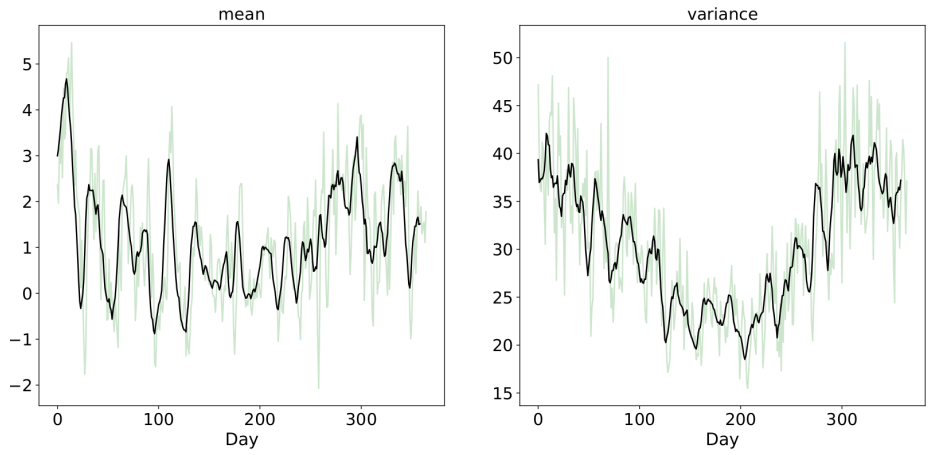
[35] Lubos Mitas and Helena Mitasova. Spatial interpolation. *Geographical information systems: principles, techniques, management and applications*, 1(2), 1999.

[36] NC Privé and RM Errico. Temporal and spatial interpolation errors of high-resolution modeled atmospheric fields. *Journal of Atmospheric and Oceanic Technology*, 33(2):303–311, 2016.

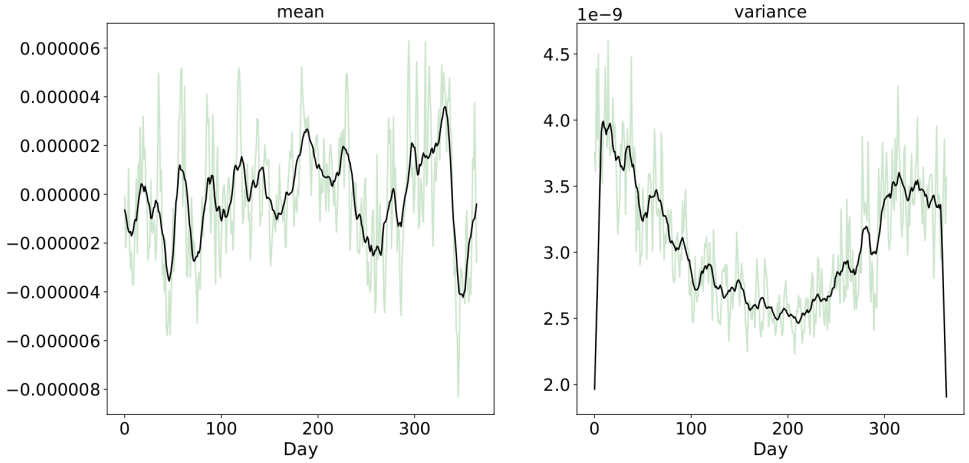
Appendix A

A.1 Seasonal Variations of the Auxiliary Variables

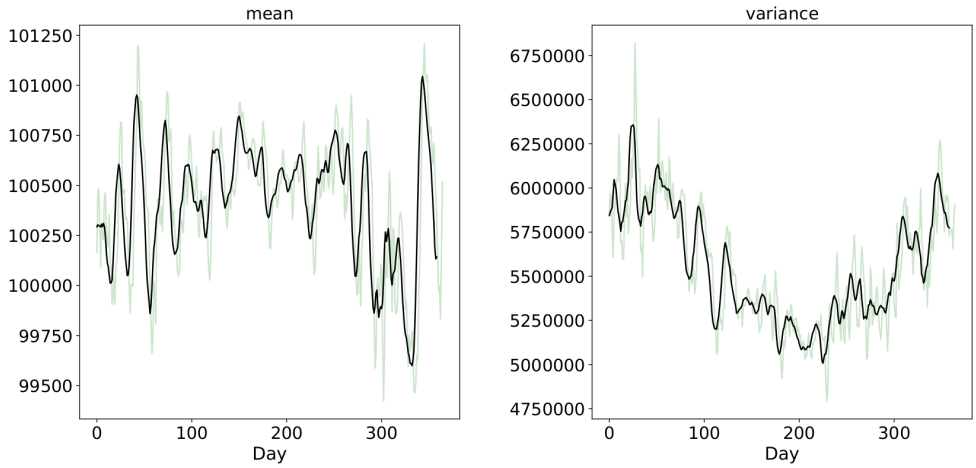
One of the reasons for including auxiliary variables in our analysis was to target the seasonality effects observed within the mean absolute errors. We show summary seasonal statistics of the auxiliary variables.



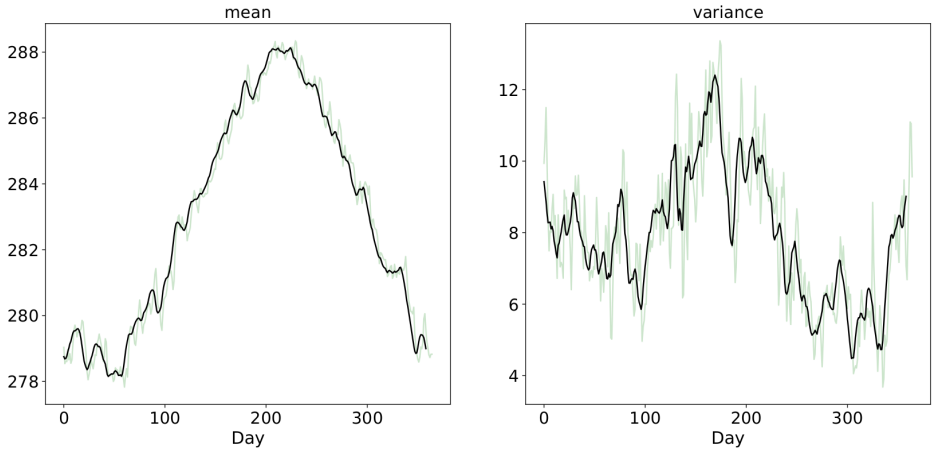
(a) v component of wind



(b) Vorticity



(c) Pressure



(d) Temperature

Figure A.0: Green shows the raw daily averaged value, and black shows a moving average of these raw values, with window size = 14, corresponding to a two week average. Note the large difference in magnitudes for the mean of these three variables, highlighting the necessity for standardisation of the features before performing linear regression.

A.2 Geographic Domains for Analysis

Figure A.1 shows the locations of the four geographic regions considered in this study. These were picked so as to represent areas with different weather patterns due to varying degrees of topographic complexity, or land-sea boundaries.

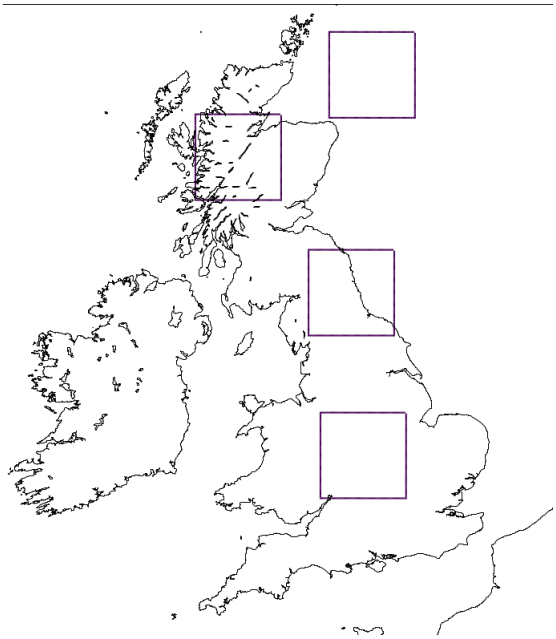


Figure A.1: Shows the British Isles and surrounding waters, where the 4 highlighted regions are the four regions considered in this study

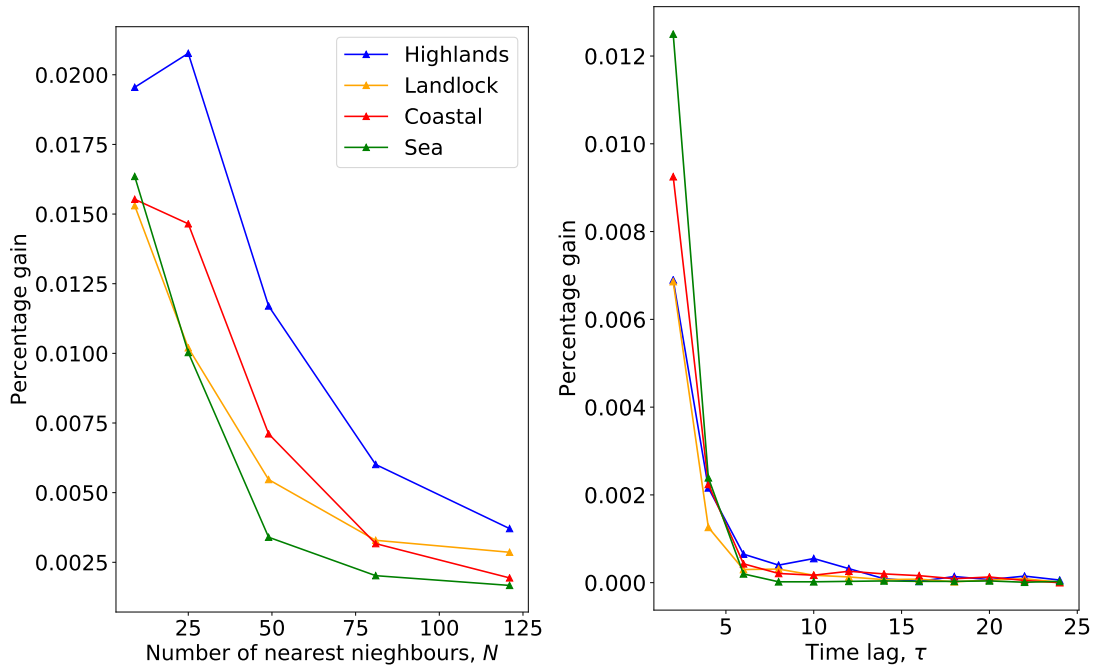
A.3 Performance Gain of Increasing the Number of Neighbours N and Time Lag τ

Here we show the performance gain when increasing the number of nearest neighbours, N, and the number of previous time points used, τ. We see that using more than τ = 5 offers almost no gain in most cases, whilst increasing N even up to 121 still offers some improvement. However, the improvement seen from using N = 49 to N = 121, is < 0.0075 in the best case, which was deemed not worth the extra computational complexity associated with the additional 72 covariates.

A.4 Evaluation of Method to Choose weight penalty λ

The method to choose the weight penalty term for both ridge and lasso regression was described in Section 5.1.2.1, however we provide a recap for the reader. We randomly selected 100 of all the possible gridpoints in a given geographical region. For these points we perform a log-space grid search between to find the optimal λ for these points. We take the mean μ_λ, of the exponent of the λ values, giving the penalty term for our models, $\lambda = 10 \times 10^{\mu_\lambda}$.

To evaluate our method for choosing the weight penalty we look at the regularisa-



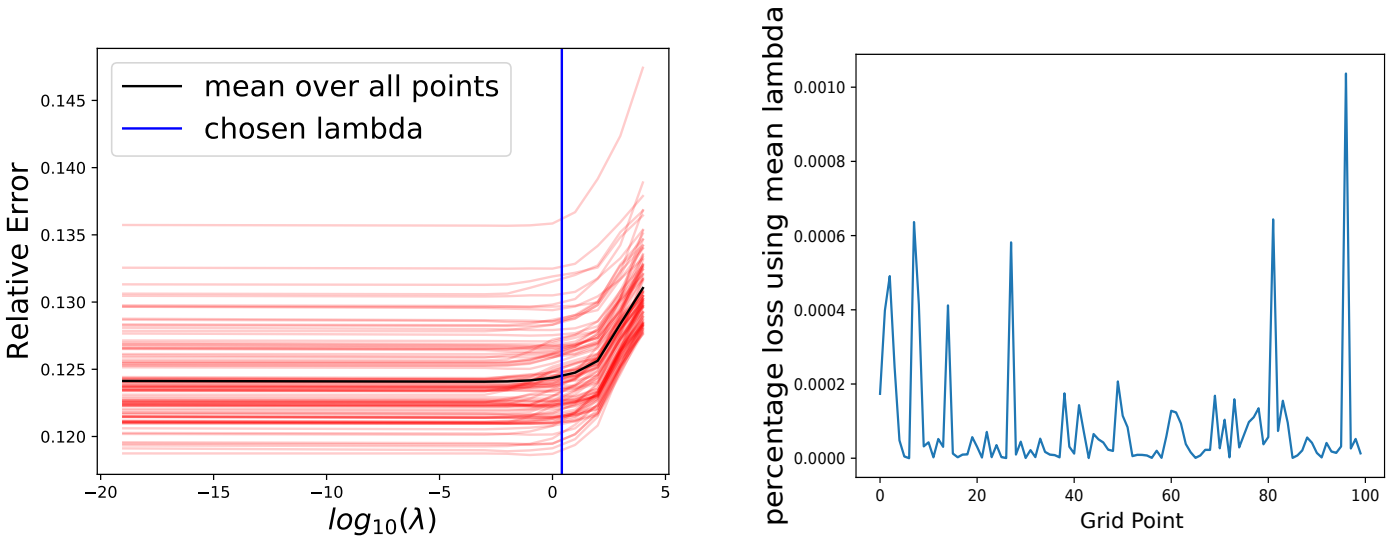
(a) (b)

Figure A.2: (a) Reduction in total relative RMSE as we vary the number of nearest neighbours ,N, (b) Reduction in total relative RMSE as we vary the time lag of previous time points used, τ.

tion path for all 100 gridpoints (Figure A.3(a)) and the losses obtained when using the chosen λ as opposed to the optimal λ for all 100 points (Figure A.3(b)). From the regularisation path, we see that very low penalties incur the same total relative RMSE. As this is the case, it is best practice to choose a higher λ value such that we learn a more parsimonious model, which is less prone to overfit to the training data [28]. We see that our method to choose λ selects a large λ that still offers low total relative RMSE. We see from A.3(b) that the loss on all 100 gridpoints doesn't rise above 0.001, further validating our approach to choosing λ.

A.5 PCA on Our Data

Here we explore principal component analysis on a sample data matrix for a given gridpoint. We expect that many of our features will be highly correlated due to their close proximity in either time or space. We see that to retain 99% of the variance we require only 49 principal components, indicating a high degree of correlation between



(a) (b)

Figure A.3: (a) Shows the regularisation path for all 100 gridpoints considered (red), where the black line shows the mean regularisation path, and blue shows our chosen lambda value. (b) Shows the differences in total relative RMSE between the predictions using the mean λ value, as opposed to the optimal λ values.

many of the features in our data matrix. This motivates the exploration of feature selection methods such as lasso regression, as discussed in the main body of this thesis.

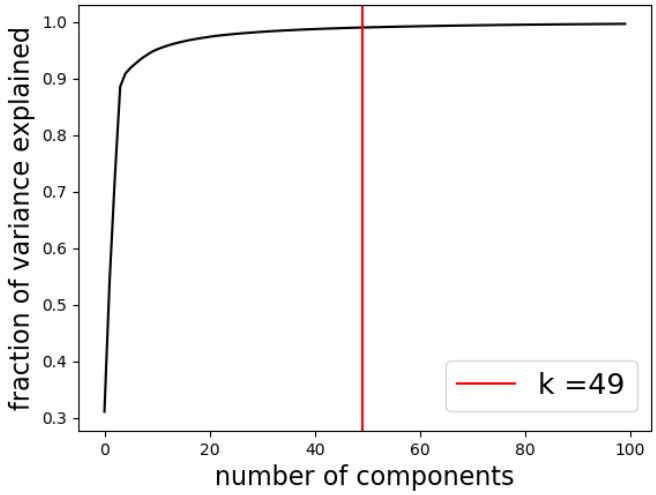


Figure A.4: Shows the fraction of variance explained as we increase the number of principal components (black). We see that using 49 PC's retains 99% of the variance (red).