INTRODUCTION:

Temperature in UK has been increasing gradually over the years. In this report we will analyse maximum temperature in UK in the year 2020. For this analysis we use two data files taken from UK Met Office. One of the datafile contain the geographic coordinates of 20 location along with its elevation (metadata). The other datafile contains Maximum temperature observed daily in these 20 locations in the year 2020 (MaxTemp). The aim of this report is to fit a suitable spatial model to predict the maximum temperature in Morecambe, Coventry, and Kinross as well as to find the effect of elevation on this temperature. We also want to fit a several time series model to predict the maximum temperature both daily and weekly for a few regions and see how well the model performs in different region.

DATA SET ANALYSIS:

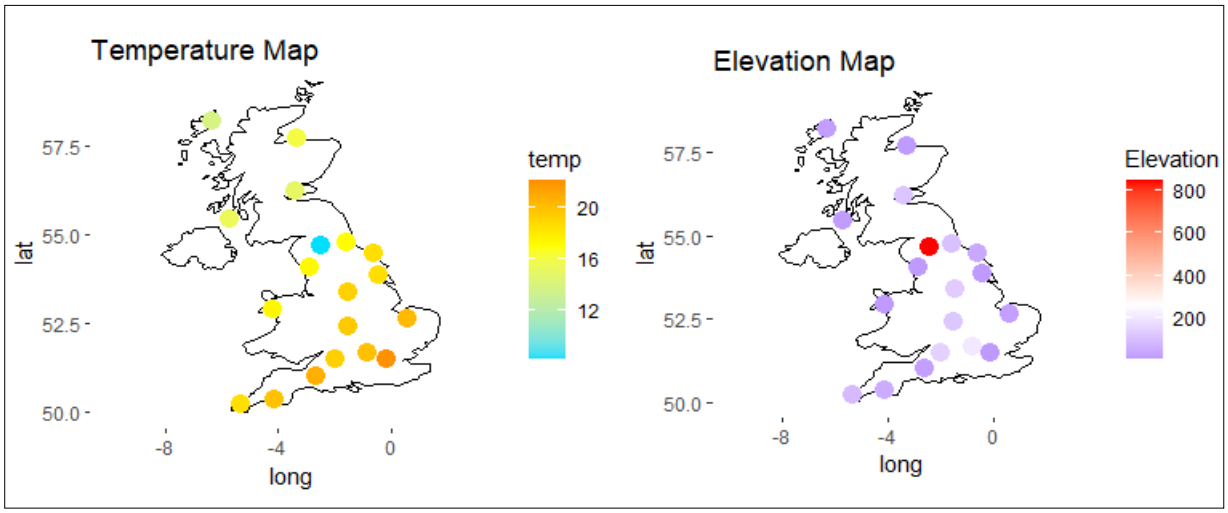

Figure 1: Temperature and Elevation map:

There are no missing values in both the dataset. From the summary of the temperature datasets (MaxTemp), we observe great deviation in temperature at the same location from the mean this is to be expected as we are taking temperature throughout the year. We also observe deviation in temperature across different regions at the same time period. This difference is understandable as the area's surveyed have different geographic location and traits. When plotting both temperature and elevation we observe regions with higher elevation have lower temperature Figure 1. To visualise the location of these 20 regions we plot a map Fig A (in appendix). The region with the highest elevation is Dun fell. Another observation that can be observed is a trend in the spread of maximum temperature spatially. In Figure 1, the temperature was measured on 12th September 2020 and on this day high temperature is observed on the southern side of UK while the northern side observes relatively lower temperature.

The time series analysis of the data provides further information on the yearly patten of temperature in the year 2020. There seems to be a yearly trend, but no cyclic pattern is observed as the variation in temperature is at random each day. Looking at Figure 2 we observe that January to April have low temperature which then gradually increases to a peak in august and then gradually decreases till December. When checking for any weekly and monthly cyclic pattern we observed none. To confirm this we plot a periodogram which confirms that there is no seasonal pattern as we do not observe a significant spike anywhere other than 0 Figure B (in appendix). The deviations observed each day from average is with some random order therefore can be considered as a white noise. To check if the time series required any transformation we check for Box Cox transformation using BoxCox.lambda and get values close to one therefore does not require transformation.

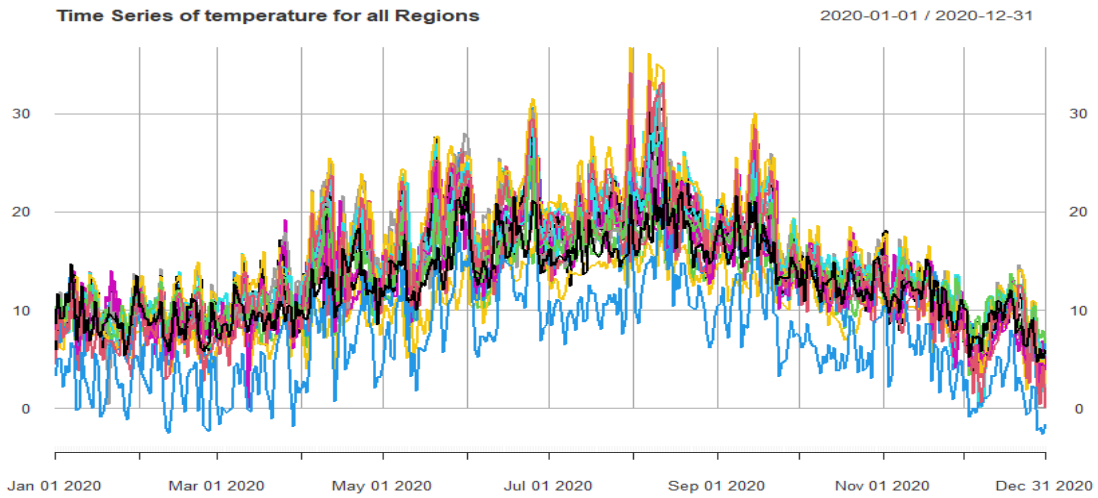

Figure 2: Timeseries for all regions

SPATIAL ANALYSIS:

Here we are going to build a spatial model to predict the Maximum temperature at Morecambe, Coventry, and Kinross on September 12th, 2020. As we have two datasets with location and maximum temperature separate, we merge the two dataset and convert it into geo data. Here we leave out the 3 locations we are predicting when creating geo data. In the geo data we provide the boundary of Britain to make prediction over a grid of Britain. When plotting the geo data Figure C (in appendix) we see a pattern as mentioned in the initial data analysis. Here we observe that high temperature clustered in the south and low temperature clustered in the north this might cause Anisotropy. We check if the data points are Anisotropic by measure the variation of temperature in different direction using variog4 function. This function shows the variance between points in different direction. From the Figure D (in appendix) we observe Anisotropy due to difference in semi variance in different direction. When we put a linear Figure E (in appendix) trend the variation in different direction becomes better therefore, we will be using a linear trend in our spatial model. If we put a polynomial trend Figure F (in appendix), we get almost the same semi variance therefore we go with a linear trend.

Here for modelling a spatial model, we have only a few locations and data points therefore we perform a Gaussian process with likelihood analysis to build a model. In gaussians approach we assume normality. Here the parameters that we estimate are mean function (β), variance (σ^2), correlation parameter (φ) and nugget.

We build multiple models and select the best-suited one by putting different covariance matrices where the parameters are estimated using maximum likelihood and restricted maximum likelihood. We use maximum likelihood instead of Bayesians as Bayesian is computationally intensive and does not perform well with a small set of data. From the summary of the models in Table 1 (in appendix) the model with the least AIC and maximum likelihood is the model with coefficient matrix 'Matern' with a linear trend and parameters estimated using restricted maximum likelihood.

To make sure the model performs well we cross-validate the model by checking its residuals. From the residuals in Figure G (in appendix) we observe that the QQ plot residual for prediction lies mostly along the 45-degree line but some points in the lower bound stray away from this line this suggests that the model does not predict lower temperature well. We also observe a relationship between fitted and true values. The plot of the coordinates X and Y shows an even spread and no trend is present. We also observe that 95% of the standard residuals lie between -2 and 2. As this is a probabilistic prediction, we observe some values outside the 2, -2 points. Here we assume normality, but the residuals suggest the

presence of non-normal trends. This might be due to the presence of very low data points used to predict a large region therefore we will continue with this model for our prediction.

From the prediction Table I, we see that the model predicts well for Kinross and Coventry but the prediction at Morecambe is off by 2 degrees this may be due to other factors like elevation not being accounted for. We also observe that the variance at these locations is high.

Region	Location (Lat, Long)	Actual value	Predicted value	Variance
Kinross	-3.413 56.214	15	15.15699	7.182458
Morecambe	-2.860 54.076	17.3	14.89081	6.291337
Coventry	-1.536 52.424	19.6	19.07928	6.510712

Table I : Actual and predicted values of Spatial model.

We visualise the model by predicting temperature at geo locations without data by plotting on a 0.1-degree grid of Britain. From the visualisation when compared with the data in Figure C (in appendix) we see that the prediction follows the pattern of the original data with higher temperature in the south and lower temperature in the north. From the variance plot, we can see that variance in regions without data points is high while variance in regions with data points is low.

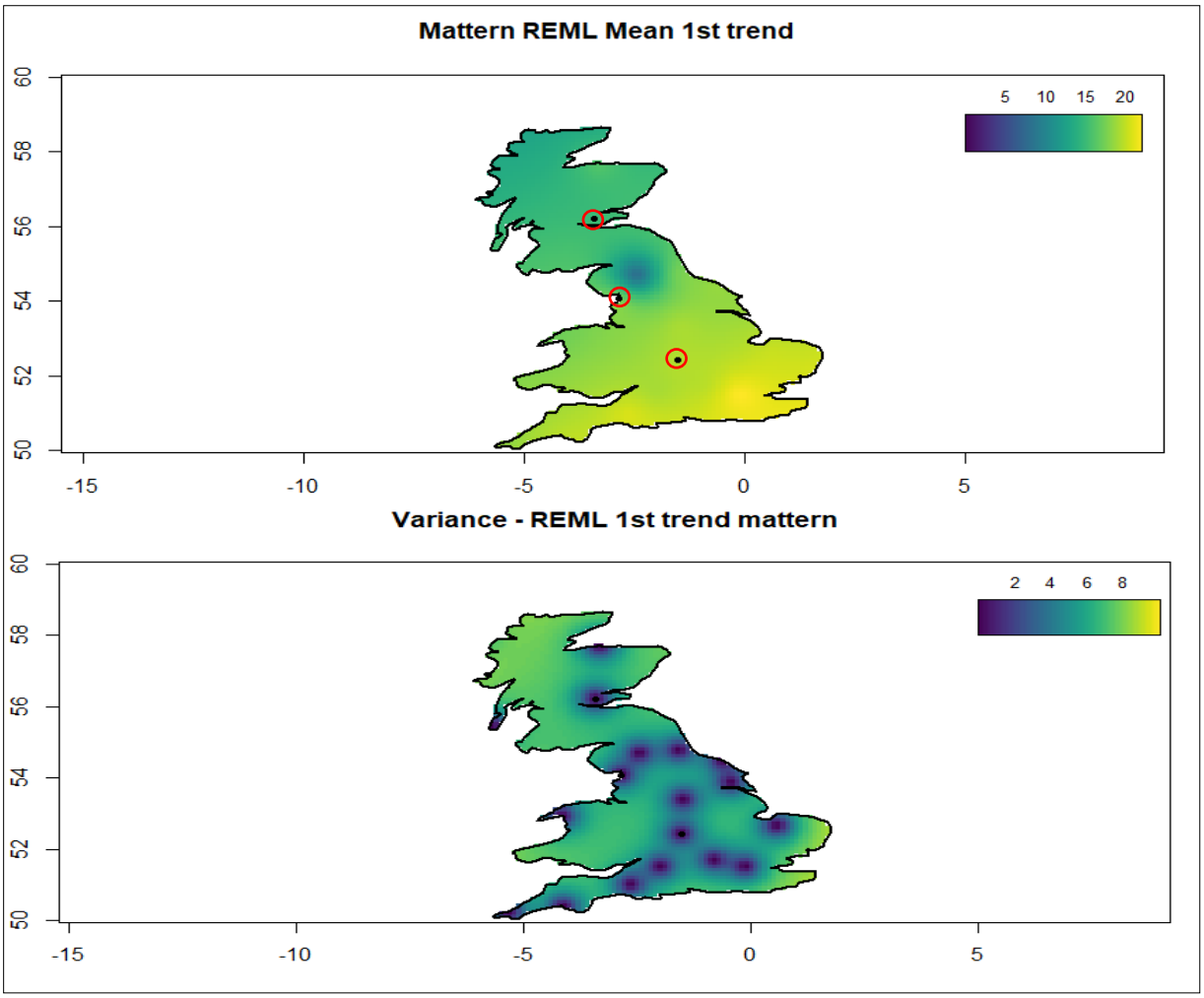

Figure 6: predicted mean and variance on UK grid

As we observe that Spatial model with location alone as explanatory variables does not predict well and has high variance we see if we can build a better model by taking other factors like elevation into account.

To build the model we first check if elevation has an effect on the model. For this we first build a simple linear model with elevation and coordinates as explanatory variables. From the summary of the model Table 2 (in appendix), we see that elevation is a significant factor as elevation has a non-zero estimate with p-values less than 0.05. Therefore, we build a spatial model with elevation. We do this using Kriging approach as we observed some non-normal trend in the Data and Kriging approach has no distribution assumption. After we fit a model, we get the predicted values as in Table II.

	Kinross	Morecambe	Coventry
Predicted Temp	15.12689	18.05174	19.48755
Variance	0.5792135	0.5744827	0.4189888

Table II: Kriging model mean and variance

Here we observe that the prediction made with the addition of the variable elevation predicts better with much lower variance.

To plot the predicted mean over the grid of UK we first create a spatial polygon from the coordinates of UK then we create a grid of 0.1 degree and crop the grid using raster's crop function. We then put the predictions of the Kriging model into the grid Figure 7.

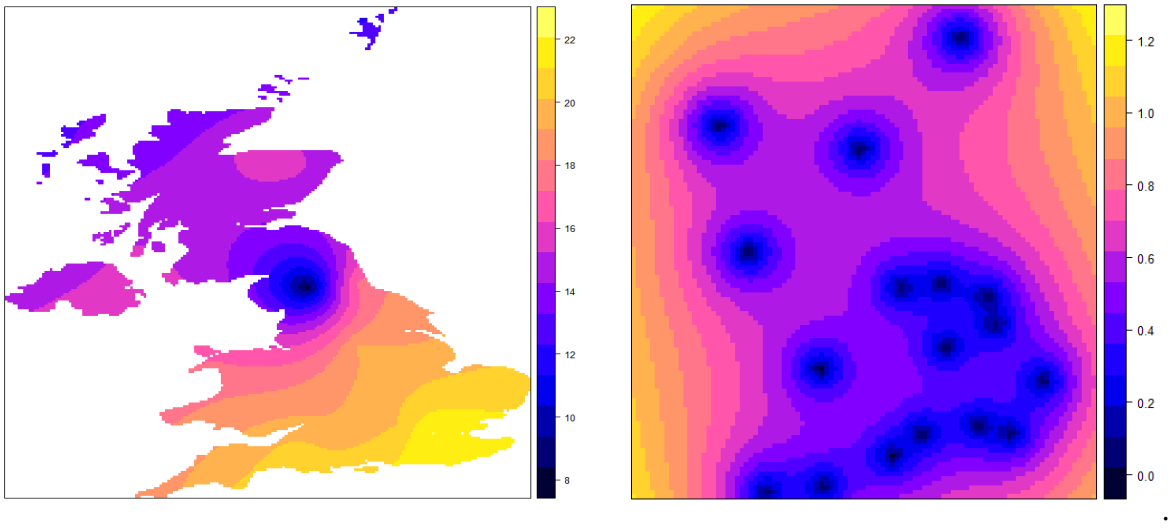

Figure 7: Mean and Variance of Kriging model plotted on grid

Here we observe the same pattern as data with higher temperatures in the south and lower temperatures in the north. From the variance plot, we observe that regions with data points have low variance while regions without data points have high variance.

TIMESERIES MODEL TO PREDICT TEMPERATURE OF YEOVILTON

Here we build a time-series model to predict maximum temperature of yeovilton on Nov 1st to 7th of 2020. For this we first convert the maximum data up to October 31st 2020 into time series Figure 8. From Figure 8 we observe that there is no cyclic pattern in the time series. But we do observe a trend with summer having higher temperature and winter having lower temperature making the time series nonstationary. Therefore, a simple ARIMA model can be used for prediction. To fit an ARIMA model we need the values of p, q and d which are the order of the auto regression, moving average and number of differencing required in the model. As there is a trend in the time series, we make it stationary by

differencing. When performing differencing once trend in the time series trend is gone and looks stationary around mean zero and further differencing does not produce much change in the trend.

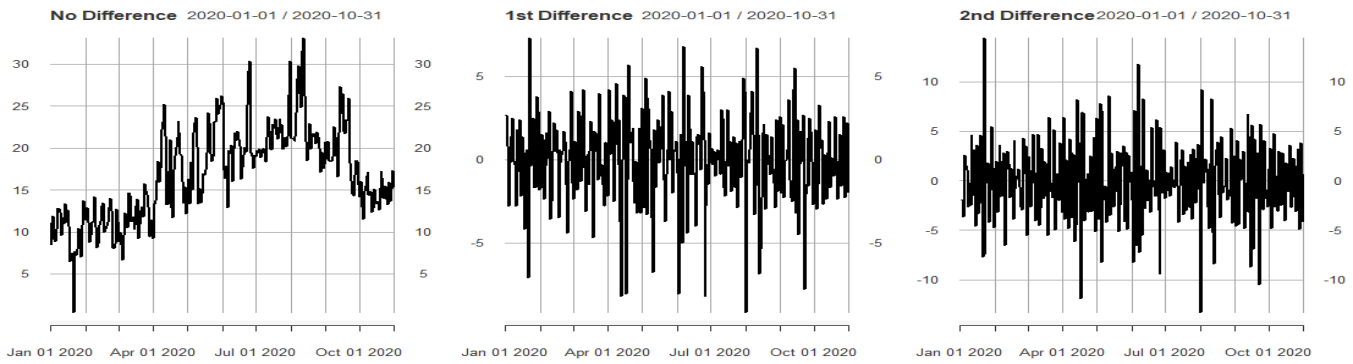

Figure 8: Time series with differencing

From the ACF and PCF Figure I (in appendix) we can find the initial values of p and q so we perform auto ARIMA. From which we get ARIMA order as p = 2, d = 1 and q = 1. The residual for this model looks good but the 95% confidence interval of the coefficient for alpha 2 contains zero this shows that 2nd coefficient for the AR part is not required. Therefore, we look for a better model. On fitting multiple models Table 3, we get ARIMA model of order (p = 1, d = 1, q = 2) as good models with low AIC, high likelihood and all the coefficients having a 95% CI without zero in it.

To cross validate the model fit we check the residuals of the model Figure H (in appendix). From the standardised Residuals we observe that majority of the residuals lie between -2 and 2. This means it is normally distributed with 95% of the points between -1.96 and 1.96. From the autocorrelation function of the residuals, we observe that there are no significant correlations at any lag other than zero. This means there seems to be no autocorrelation in the residuals. To check if the correlation at all lag is 0 we set the null hypothesis that correlation at lag k = 0 and check the pvalue. Here we find that p-value at all points is greater than 0.05 thereby failing to reject the null hypothesis. In conclusion, the model has good residuals and they look like white noise.

As we found a suitable model we use this to predict the max temperature of Yeovilton between 2020 Nov 1st and 7th and plot the prediction Figure 9.

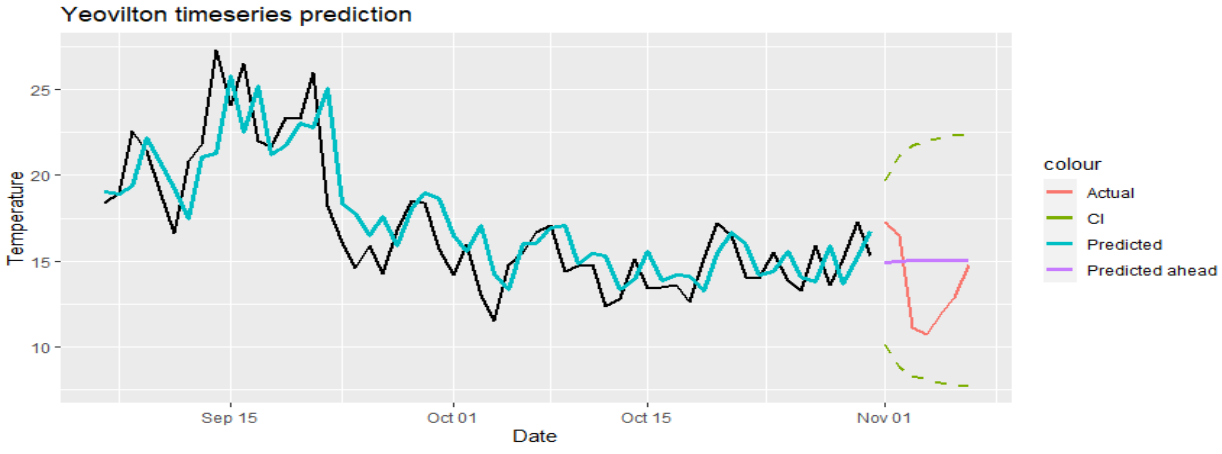

Figure 9: Yeovilton temperature prediction between 2020 Nov 1st and 7th

Table 4 in appendix gives us the predicted values. The blue line in Figure 9 is the temperature predicted by the model where data is present. Here we observe that the model picks up most of the variations of the actual timeseries. We extend this prediction by predicting 7 days ahead from Nov 1st to 7th. We observe the prediction as a straight line with a very subtle increase in temperature. The prediction looks good as the confidence interval (green dotted line) of the prediction contains the actual value that's predicted, but we observe that the CI is high therefore there is a large uncertainty in the prediction. Overall, the model predicts well.

Now we want to check how well this ARIMA model fit other regions. To check the performance of the model we fit the model to two regions where there is largest deviation from the temperature of Yeovilton as in the case of London and Dun fell and one location far away from Yeovilton, as in the case of Stornoway, and when one location is close to Yeovilton, as in the case of Lyneham.

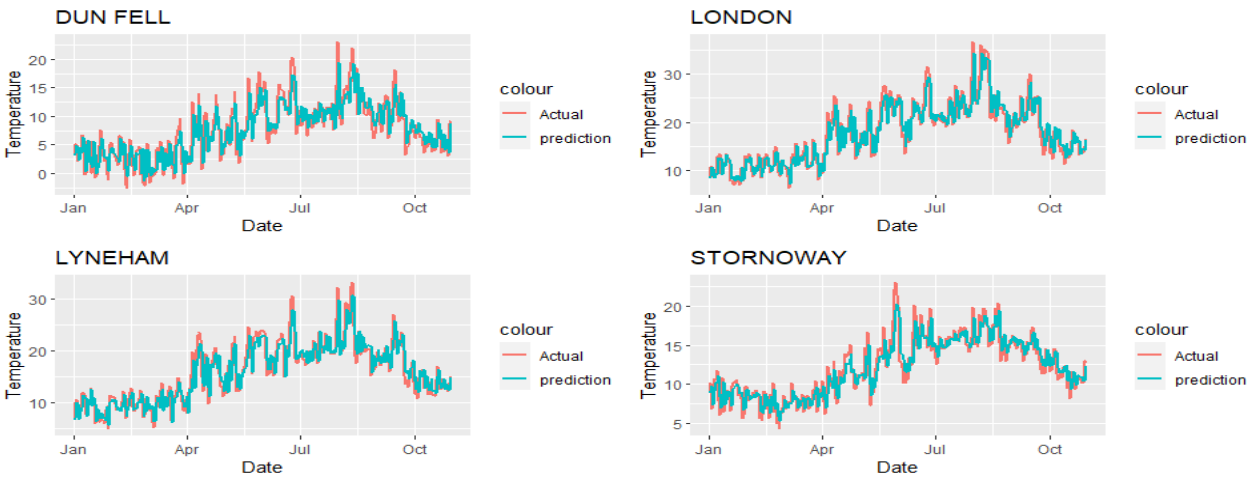

Figure 10: predictions made using ARIMA model (1,1,2)

The predictions made by the ARIMA model (1,1,2) over the region Dun fell, London and Lyneham are good Figure 10. The predicted coefficients have a 95% CI without zero in it and the standardised residuals of these regions also have normal distribution with no significant correlation at any lag other than zero Therefore behaving like white noise Figure J to L (in appendix). As the model predicts well and the residuals look good the model is a good fit for these locations.

But for the region Stornoway we observe many of the higher value predictions are greater deviation than the actual value Figure 10. The CI of the predicted coefficients in Table 5 (in Appendix) for beta 2 has zero therefore this coefficient seems insignificant. The residuals of the model Figure M (in appendix) also do not look good with moth of the p- values at lag 't' lower than 0.05 thus rejecting the null hypothesis that correlation at lag k = 0. From this we conclude the model predicts well for nearby location but does not predict well for Stornoway which is further away. This maybe due to large distance away from Yeovilton causing large difference in temperature during the same date.

We can use this data set to check how well a weekly average model predicts when compared to a model with daily data. For this analysis, we will take only Yeovilton as a location and predict 3 weeks from week 51 to week 53 using both daily and weekly averaged model and compare them. Here we first create a data set with weekly averaged maximum temperature. Then we convert the data set into timeseries, and we get the plot as in Figure N (in appendix). From the figure N (in appendix), we observe

that the time series is not stationary and requires differencing we also observe that there is no seasonal pattern and the variations of temperature each day is random but follows a trend of low temperature from week 1 to week 14 and higher from 15 week to 40 then again lower temperature after week 40. To make sure there is no seasonality we plot a periodogram Figure O (in appendix). From the periodogram we do not observe any seasonality therefore we use ARIMA model to fit a time series.

By using the diff function, we see that differencing of 1 works well. To get a rough idea of what kind of ARIMA model to fit we perform auto ARIMA here we get ARIMA with order (1,1,0) as an initial good fit. To check if there is a better model, we perform manual fitting by changing p and q values. From the summary of all these models in Table 6 (in appendix) we get that ARIMA model of order (1,1,0) is the best model with low AIC and high Likelihood. The CI of the coefficients of the model also does not contain zero. To cross validate we check the residuals Figure P (in appendix). The residuals look good with no significant correlations at any lag other than zero and the residuals normally distributed with 95% of the points between -1.96 and 1.96.

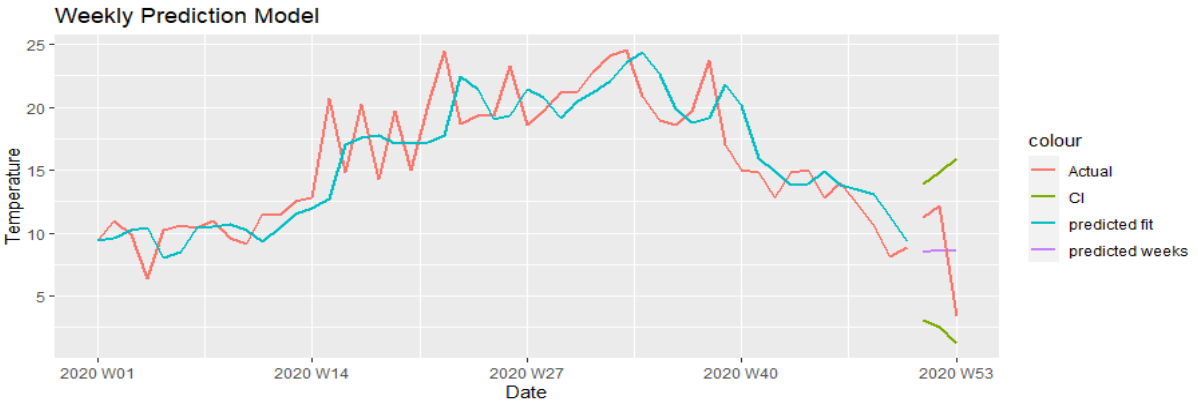

Figure 11: Weekly ARIMA model prediction

From Figure 11 we see that the model does not fit the data properly. The blue line is the predicted fit and does not fit the actual value red line properly. The violet line is the extended 3-week prediction and the confidence interval of the extended 3 weeks contains the actual value.

To compare this with daily data we fitted a ARIMA model to a daily time series. By fit different models similar to how we fit the other ARIMA models we found ARIMA (1,1,2) as the best model from summary Table 7 (in appendix) and residuals Fig Q (in appendix) looks good with no significant correlations at any lag other than zero and the residuals normally distributed with 95% of the points between -1.96 and 1.96.

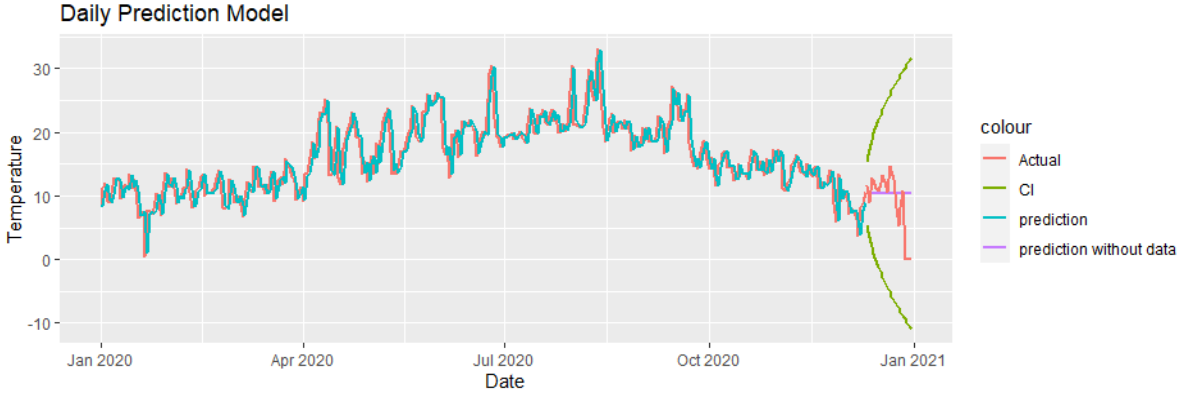

Figure 12: Daily ARIMA model prediction

From Figure 12 we observe that the CI is quite large compared to the CI of weekly prediction this might be because we are predicting 21 data points in the daily model while only predicting 3 points ahead in the weekly model. The confidence interval here contains the actual values which is good. To compare the predictions, we fist calculate the weekly average of the daily prediction. After averaging we compare both the prediction to the actual value Table III.

Week	Actual model	Daily model averaged	Weekly model
W 51	11.3	10.381	8.52
W 52	12.21	10.383	8.67
W 53	3.4	10.383	8.6

Table III: Prediction and actual value of weekly averaged models.

The predictions made by both models have large deviation from the actual value for the 53rd week. Both predictions are accurate but relatively the averaged daily model is better for the week 51 and 52 and weekly model is closer to the actual value for week 53.

SUMMARY:

From the spatial analysis we can summarise that additional explanatory variables will improve the model prediction and reduce the variance. From the modelling we also found that performing a linear trend when Anisotropic condition or trend is present produces better prediction. We also observe that assuming normality even though the data is not normal reduces the accuracy of the model. When the model is not normal Kriging approach suits the analysis. We also observe that variance of regions without data points increases with increase in distance away from region with data point.

From the Timeseries analysis we observed that ARIMA model performs well when we fit it to a model without seasonality. We observed that with increasing distance from a location we can't use the same model as the temperatures at the same period of time will be much more different. We also observed that when using averaged weekly time series we lose the prediction accuracy up to a point and then accuracy of both models decrease. When comparing the CI the CI of daily model is larger than that of weekly model.

APPENDIX:

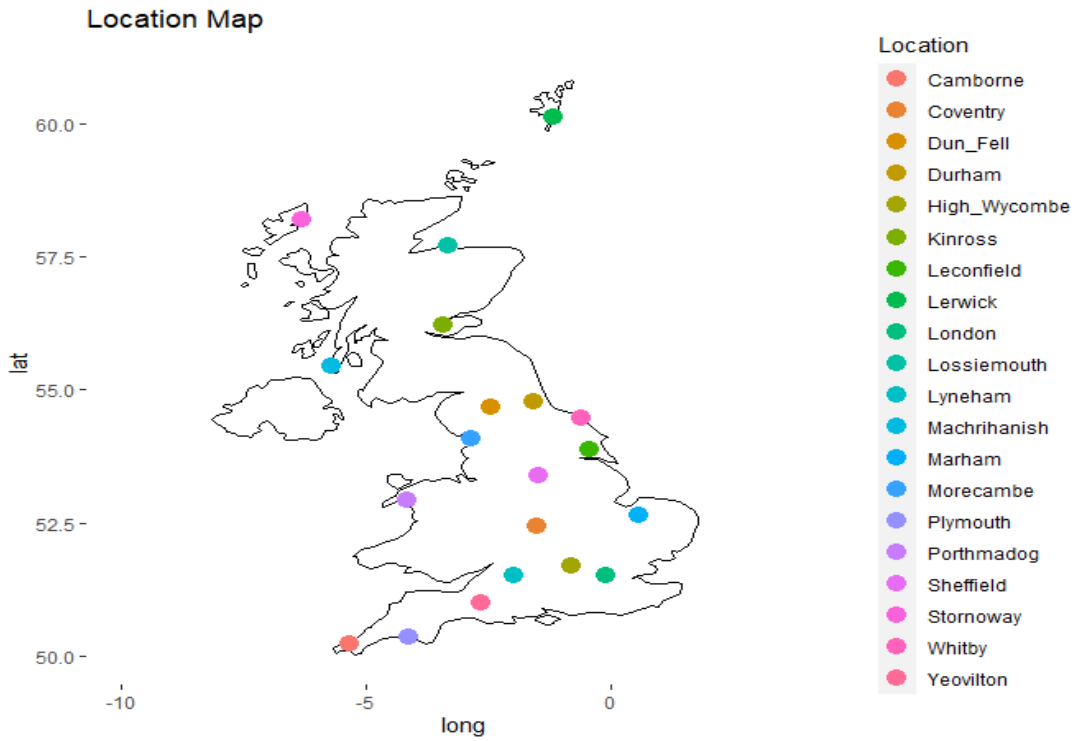

Figure A: location of all regions

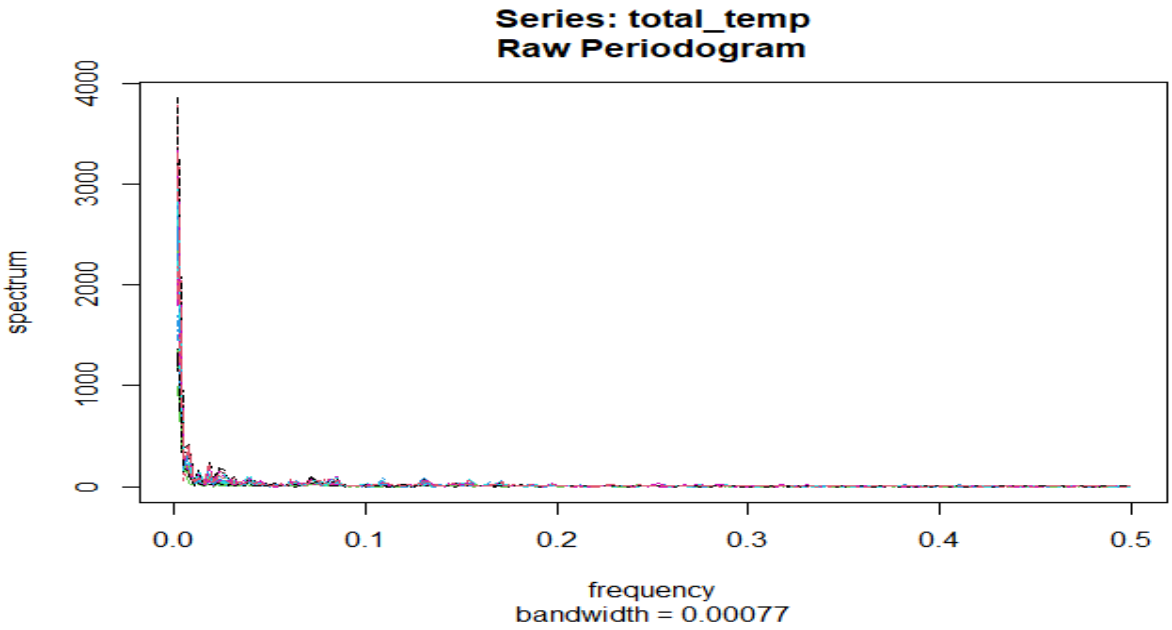

Figure B: Periodogram of daily timeseries for all regions

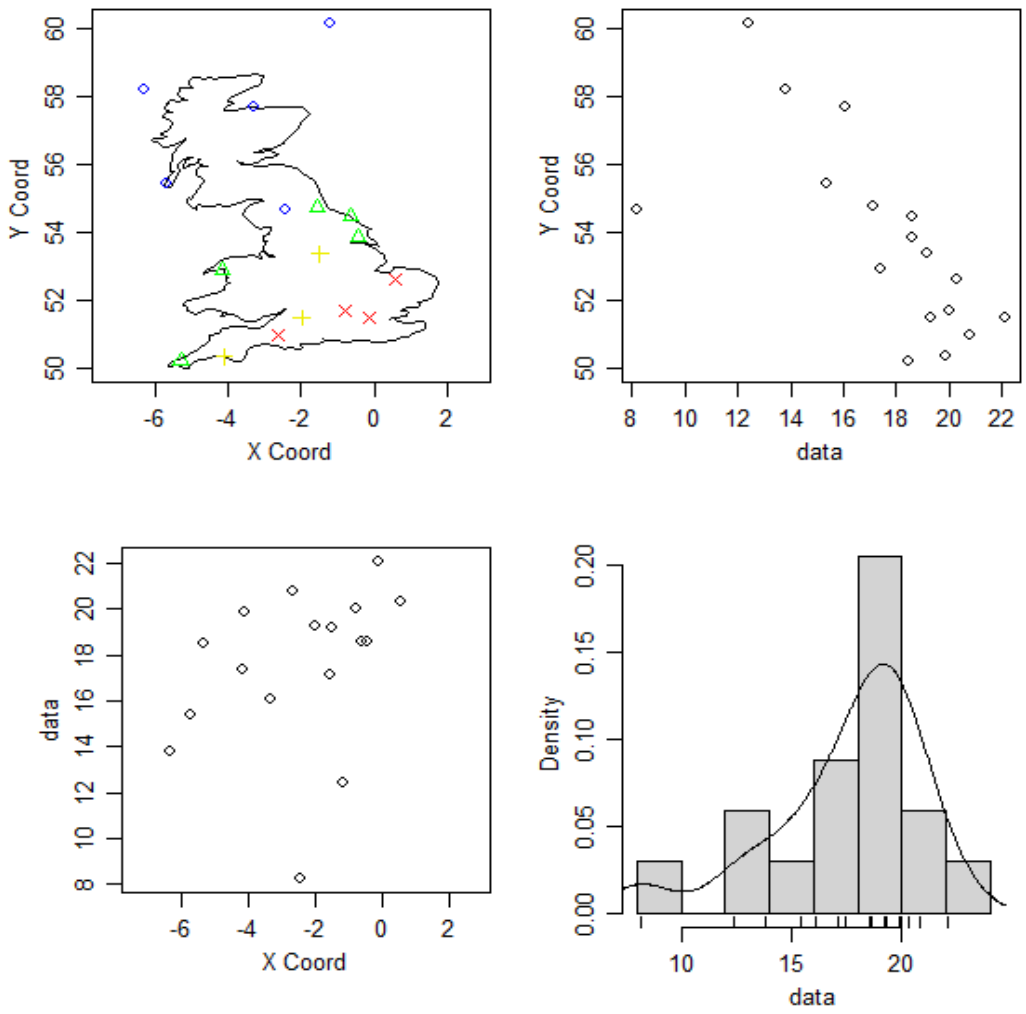

Figure C: Geo Data plot

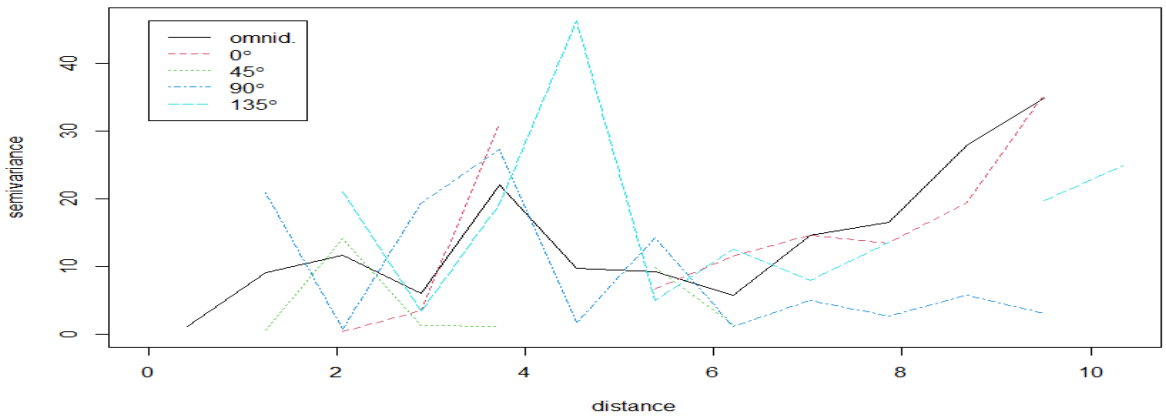

Figure D: Variog4 without trend.

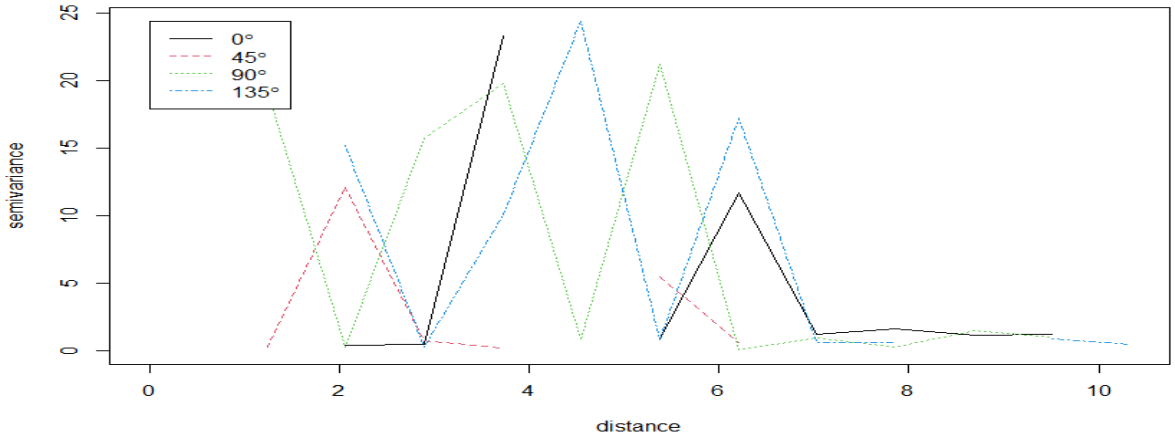

Figure E: Variog4 with 1st trend.

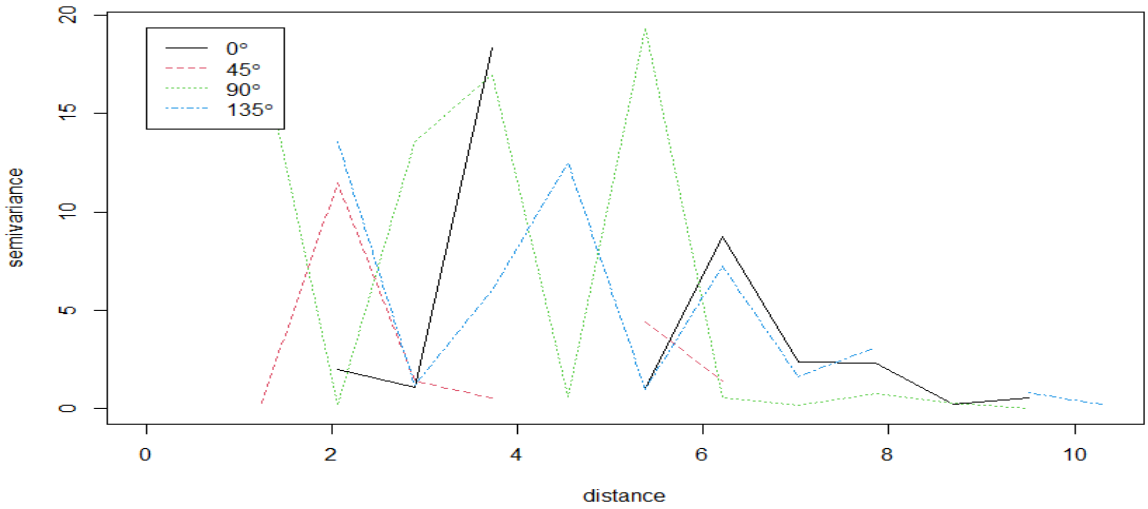

Figure F: Variog4 with 2nd trend.

Model	Log Likelihood	AIC	BIC
MATERN/ML/NO TREND	-43.04	94.08	97.41
MATERN/REML/NO TREND	-39.33	86.67	90
MATERN/ML/ LINEAR TREND	-38.12	88.24	93.24
MATERN/REML/ LINEAR TREND	-32.65	77.3	82.3
EXPONENTIAL/ML/NO TREND	-43.06	94.12	97.46
EXPONENTIAL/REML/NO TREND	-39.58	89.34	90.5
EXPONENTIAL/ML/ LINEAR TREND	-38.12	88.25	93.25
EXPONENTIAL/REML/ LINEAR TREND	-32.66	77.32	82.32
POWERED EXPONENTIAL/ML/NO TREND	-43.03	94.05	97.38
POWERED EXPONENTIAL/REML/NO TREND	-39.41	86.82	90.15
POWERED EXPONENTIAL/ML/ LINEAR TREND	-38.12	88.23	93.23
POWERED EXPONENTIAL/REML/ LINEAR TREND	-32.65	77.3	82.3

Table 1: Summary of different spatial models.

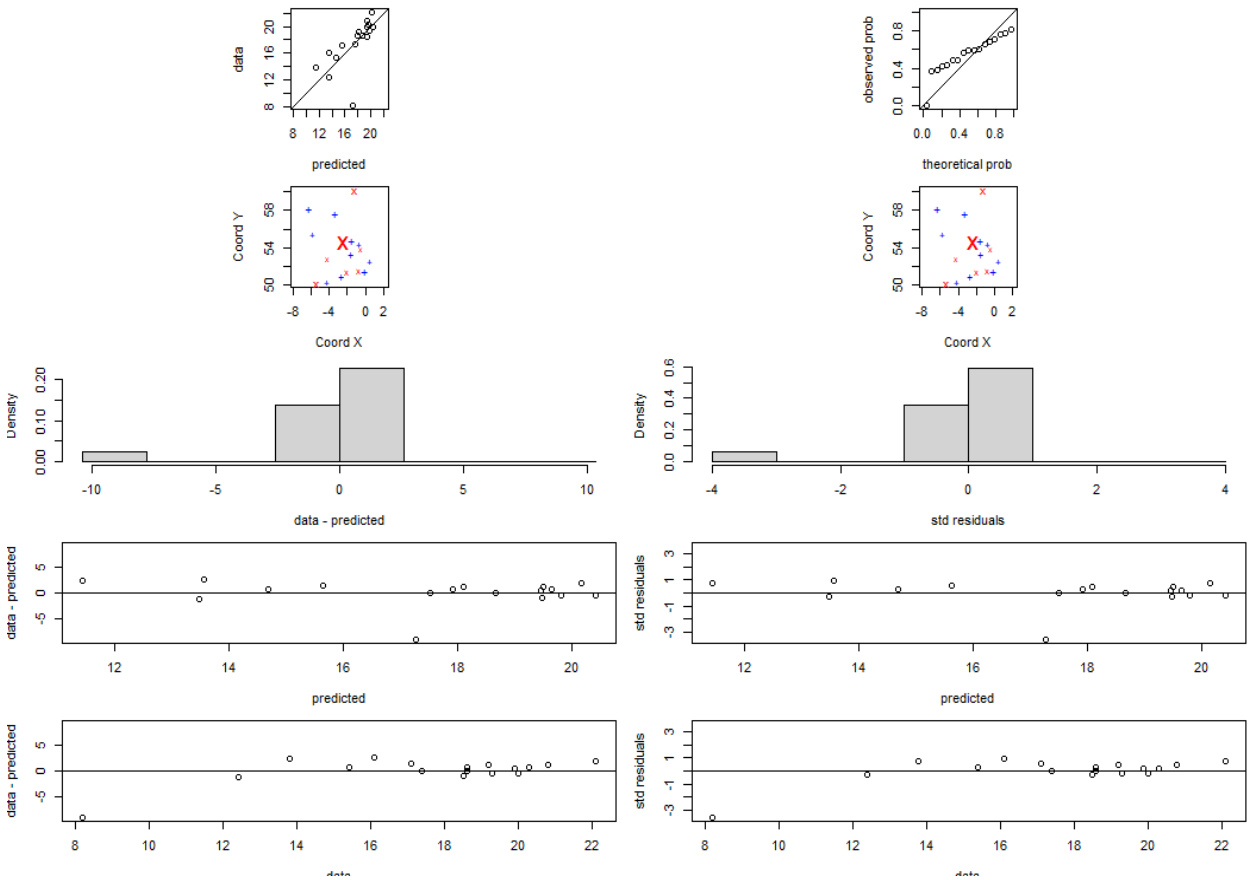

Figure G: Residual plot for spatial model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	62.822799	3.789456	16.578	3.99e-10 ***
long	0.500169	0.097679	5.121	0.000197 ***
lat	-0.797647	0.070749	-11.274	4.40e-08 ***
elv	-0.011198	0.001001	-11.182	4.85e-08 ***

Table 2: summary of linear model

MODEL NAME	ORDER (p, d, q)	LIKELIHOOD	AIC	VARIANCE
model_arima2	(1,1,1)	-704.45	1414.9	6.017
model_arima3	(1,1,2)	-702.35	1412.7	5.933
model_arima4	(1,1,3)	-702.35	1414.7	5.933
model_arima5	(0,1,1)	-720.48	1444.95	6.7
model_arima6	(0,1,2)	-710.5	1427.01	6.268
model_arima7	(2,1,0)	-714.77	1435.54	6.451

Table 3: ARIMA model summary:

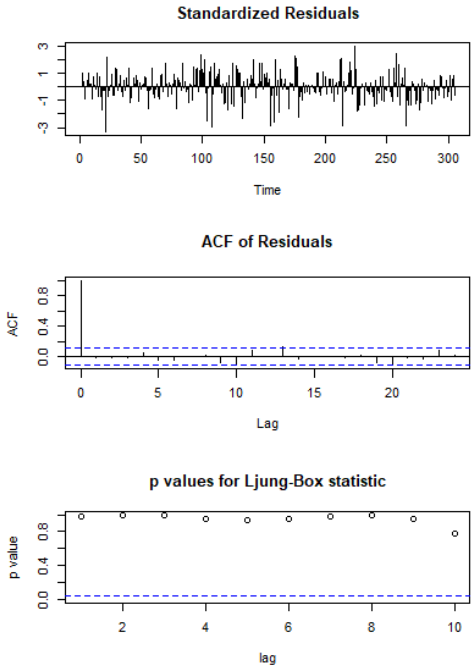

Figure H: residual of ARIMA (1,1,3)

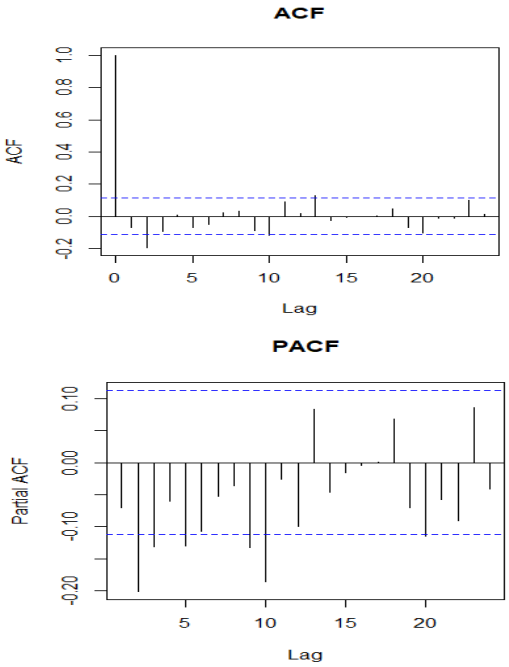

Figure I: ACF and PACF of Yeovilton 1st trend TS

Date	Actual	Predicted	Lower CI	Upper CI
2020-11-01	17.3	14.93705	10.162775	19.71133
2020-11-02	16.5	14.99371	8.832735	21.15469
2020-11-03	11.1	15.02427	8.337138	21.71140
2020-11-04	10.7	15.04075	8.088357	21.99315
2020-11-05	12	15.04964	7.931801	22.16748
2020-11-06	12.9	15.05443	7.814734	22.29413
2020-11-07	14.8	15.05702	7.716362	22.39768

Table 4: ARIMA (1,1,2) prediction from Nov 1st to 7th of 2020.

STRONOWAY	VALUE	C.I
ALPHA	0.59	(0.41,0.77)
BETA1	-0.906	(-1.11, -0.69)
BETA2	-0.0045	(-0.17,0.164)

Table 5: CI and coefficients of ARIMA model in Stornoway

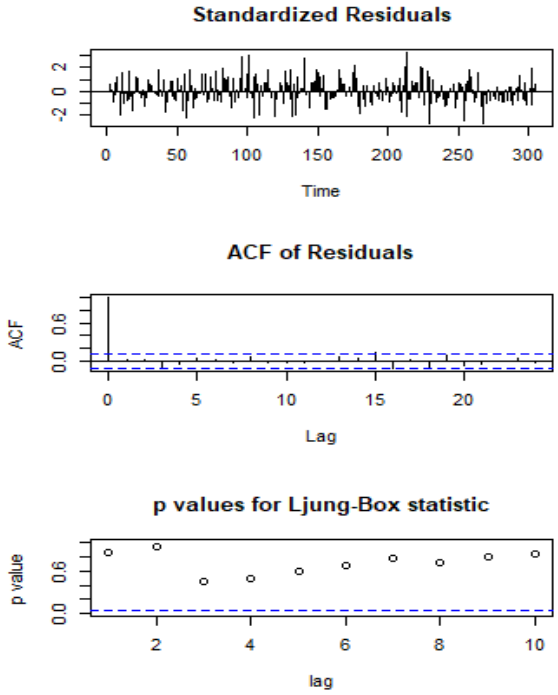

Figure J: ARIMA Dun fell residual

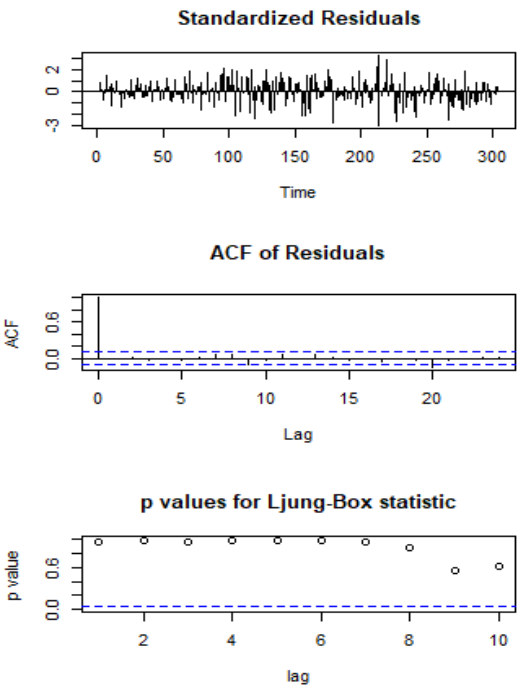

Figure K: ARIMA London residual

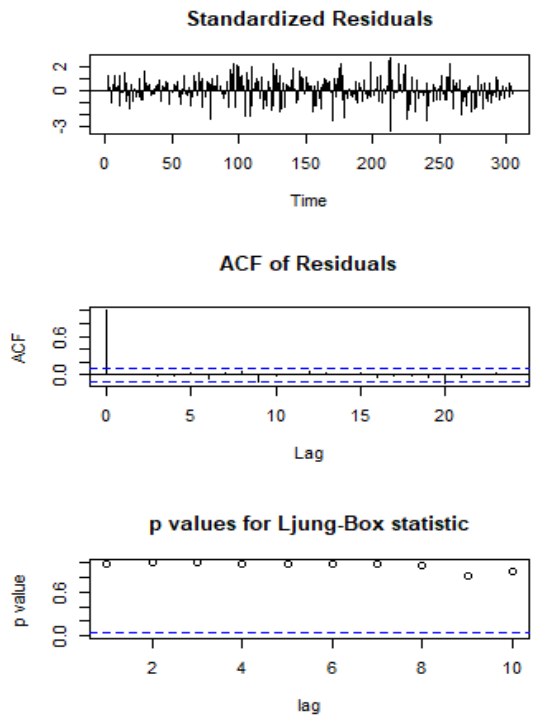

Figure L: ARIMA Lyeham residual

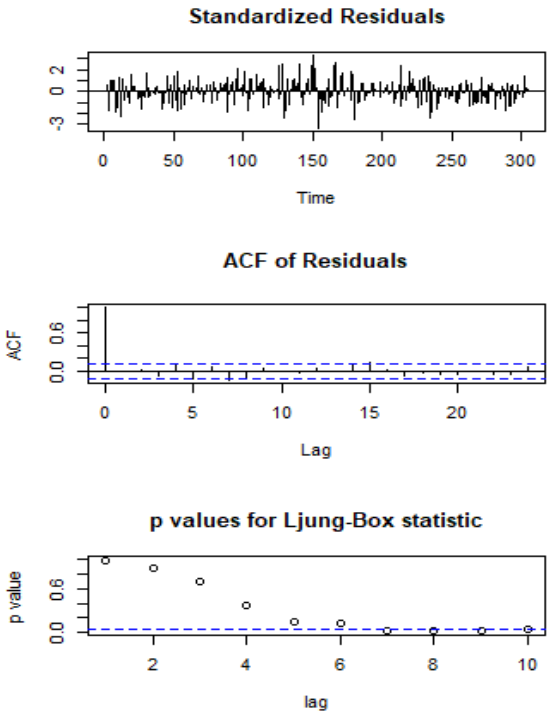

Figure M: ARIMA Stornoway residual

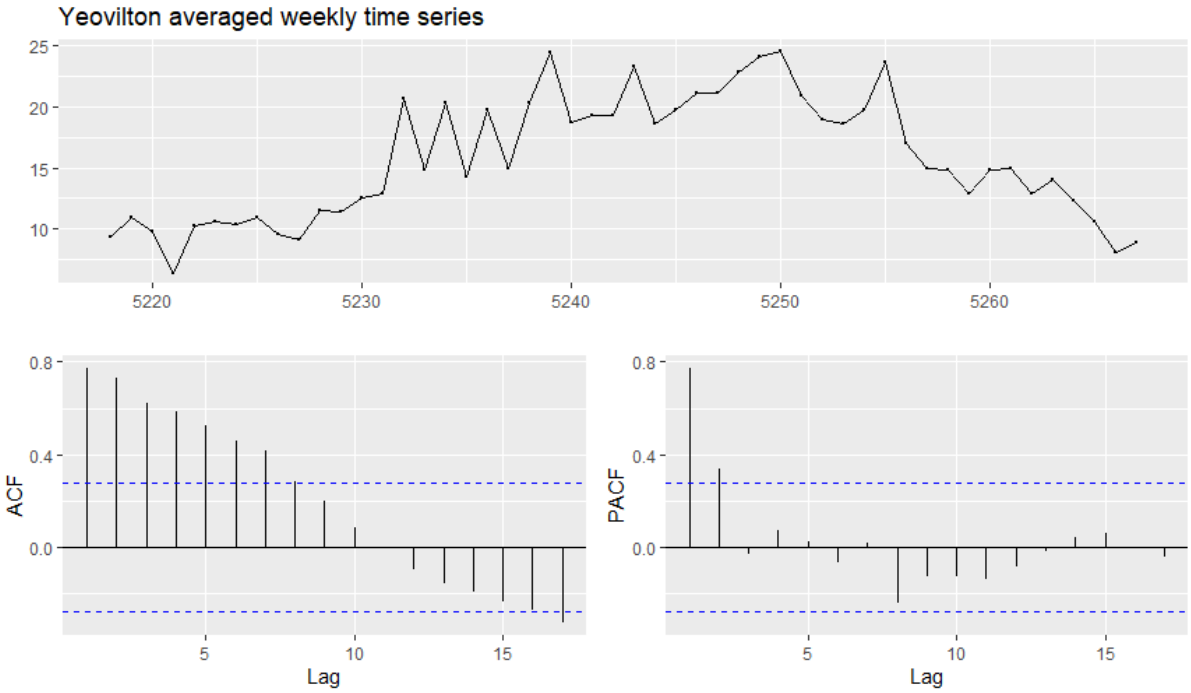

Figure N: Yeovilton weekly averaged maximum temperature time series.

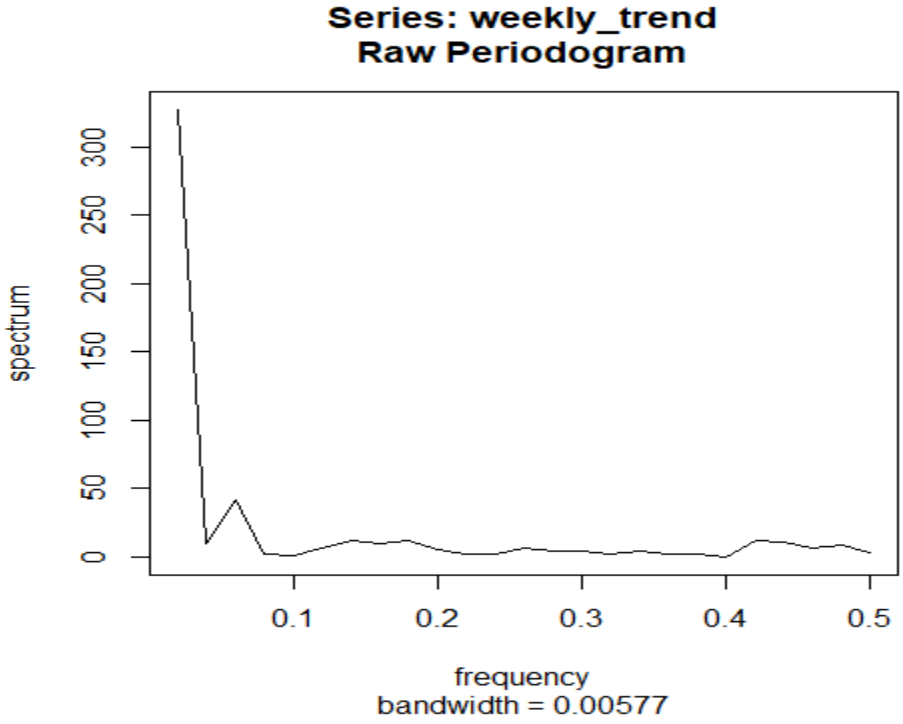

Figure O: Weekly periodogram

MODEL NAME	ORDER (p, d, q)	LIKELIHOOD	AIC
model_ARIMA0	(0,1,1)	-119.94	243.88
model_ARIMA1	(0,1,2)	-119.64	245.27
model_ARIMA2	(1,1,0)	-119.57	243.14
model_ARIMA3	(2,1,0)	-119.54	245.07
model_ARIMA4	(1,1,1)	-119.48	244.97
model_ARIMA5	(1,1,2)	-119.21	246.43

Table 6: ARIMA model summary for weekly averaged time series.

MODEL NAME	ORDER (p, d, q)	LIKELIHOOD	AIC
model_ARIMA0_d21	(0,1,1)	- 810.68	1625.36
model_ARIMA1_d21	(0,1,2)	799.08	1604.17
model_ARIMA2_d21	(1,1,0)	811.57	1627.15
model_ARIMA3_d21	(2,1,0)	804.59	1615.18
model_ARIMA4_d21	(1,1,1)	792.98	1591.95
model_ARIMA5_d21	(1,1,2)	790.97	1589.94
model_ARIMA6_d21	(2,1,1)	791.09	1590.18

Table 7: ARIMA model summary for daily time series.

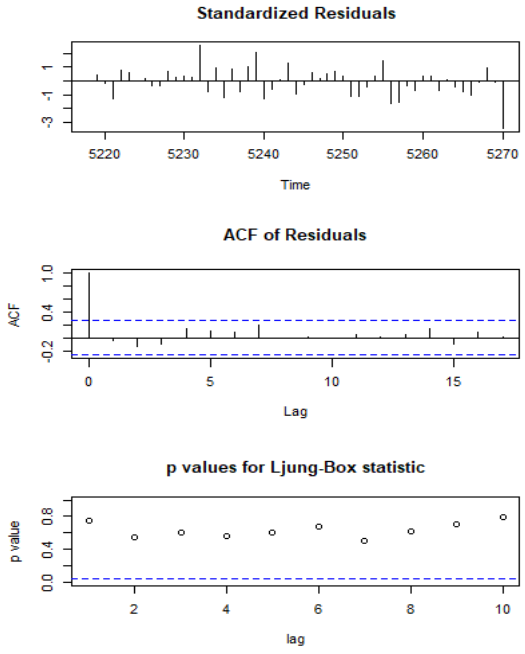

Figure P: Residual of Weekly model

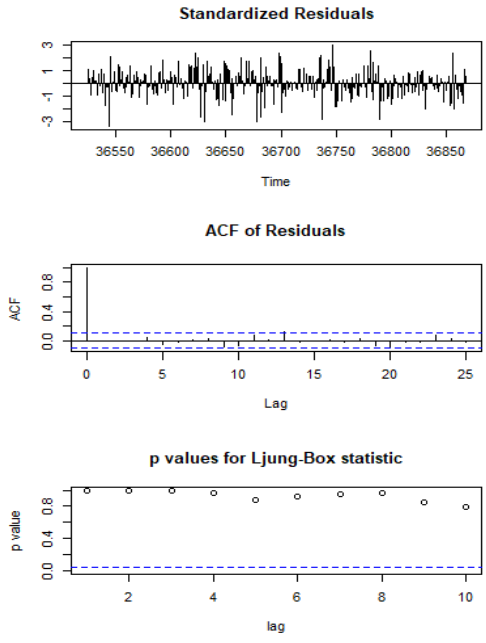

Figure Q: Residual of Daily model