Spatial Interpolation of Temperature in Southwest Nova Scotia

MScAG Project by Dana Melamed

May 31, 2018

Acadia University and The Centre Of Geographic Sciences

Contents

[List of Figures ii](#_Toc520026162)

[List of Tables iii](#_Toc520026163)

[1. Introduction 1](#_Toc520026164)

[Study Area 4](#_Toc520026165)

[2. Methods and Data 6](#_Toc520026166)

[Data 6](#_Toc520026167)

[Method Stages Overview 11](#_Toc520026168)

[Stage 1: Prepare Interpolation Inputs 12](#_Toc520026169)

[Raster Brick 12](#_Toc520026170)

[Data Frame 13](#_Toc520026171)

[Separate Model and Validation Stations 14](#_Toc520026172)

[Stage 2: Model Selection 15](#_Toc520026173)

[Variability Coverage by Weather Stations 15](#_Toc520026174)

[Timeframe 16](#_Toc520026175)

[Stage 3: Generating Output 17](#_Toc520026176)

[3. Results and Discussion 18](#_Toc520026177)

[Model Selection 18](#_Toc520026178)

[Timeframe 18](#_Toc520026179)

[Daily Temperature Variable 19](#_Toc520026180)

[Knots in smooth terms 19](#_Toc520026181)

[Accumulated Monthly GDD 21](#_Toc520026182)

[4. Appendix A: Sample Scripts for Data Preparation 25](#_Toc520026183)

[1. Modify Aspect and Proximity to the Coastline Rasters (R) 25](#_Toc520026184)

[2. Resample Solar Radiation Rasters (Python) 25](#_Toc520026185)

[3. Create Raster Brick 25](#_Toc520026186)

[4. Prepare Data Frame 26](#_Toc520026187)

[Appendix B: Plots used in model selection 26](#_Toc520026188)

# List of Figures

[Figure 1‑1 Weather stations locations throughout SWNS. Origins are shown by colour: AGRG (grey), DNR (blue), and ECAN (red). Stations are labelled by their unique station I.D. 3](#_Toc519356926)

[Figure 1‑2: Illustration of how weather station data is modelled with rasters. a) Weather stations at corresponding cells of DEM. b) Temperature records plotting against DEM value with a linear regression line 0](#_Toc519356927)

[Figure 1‑3 Study area, Southwest Nova Scotia, Canada 1](#_Toc519356928)

[Figure 1‑4 Relative fruit and tree nut farm in Canadian Provinces 2016 (Stats Canada) 0](#_Toc519356929)

[Figure 1‑5 Relative nursery and tree production in Canadian Provinces 2016 (Stats Canada) 0](#_Toc519356930)

[Figure 1‑6 The four major wine producing regions in Nova Scotia 0](#_Toc519356931)

[Figure 2‑1 The fourteen primary watersheds in SWNS that form the boundary of the study area. 1](#_Toc519356932)

[Figure 2‑2 The DEM of SWNS used in analysis. Reprojected, cropped and resampled from Nova Scotia Department of Natural Resources. 1](https://d.docs.live.net/fb6e9bcbbea73c46/MSc/Report/FirstDraft_2.docx#_Toc519356933)

[Figure 2‑3 The aspect raster in used in analysis. Generated with the DEM in 1](https://d.docs.live.net/fb6e9bcbbea73c46/MSc/Report/FirstDraft_2.docx#_Toc519356934)

[Figure 2‑4 Flowchart of the methods 1](#_Toc519356935)

[Figure 2‑5 The coverage of inputl raster values by weather stations. 1](#_Toc519356936)

[Figure 3‑1 Histograms of errors for the three daily temperature variables modelled at five timeframes 4](#_Toc519356937)

[Figure 3‑2 Histogram of errors at validation stations from daily mean temperature models at three different timeframes (daily, weekly, month) with three different values of k in the smooth terms (1, 5, 9, no limit). 6](#_Toc519356938)

[Figure 4‑1 11](#_Toc519356939)

[Figure 4‑2 Distribution of residuals relative to three units of daily temperature (minimum, maxmimum and mean) at five different timeframes (all years, yearly, monthly, weekly, and daily) 11](#_Toc519356940)

[Figure 4‑3 GCV scores from modelling with three units of daily temperature (minimum, maximum and mean) at five different timeframes (all years, yearly, monthly, weekly, and daily) 11](#_Toc519356941)

[Figure 4‑4 Adjusted R2 values from modelling with three units of daily temperature (minimum, maximum and mean) at five different timeframes (all years, yearly, monthly, weekly, and daily) 12](#_Toc519356942)

[Figure 4‑5 P-Values of smooth terms from GAMs of daily temperature variables modelled daily against raster variables individually with the formula "temp variable ~ s(raster variable)" 12](#_Toc519356943)

[Figure 4‑6 P-Values of smooth terms from GAMs of daily temperature variables modelled weekly against raster variables individually with the formula "temp variable ~ s(raster variable, yday)" 13](#_Toc519356944)

[Figure 4‑7 P-Values of smooth terms from GAMs of daily temperature variables modelled monthly against raster variables individually with the formula "temp variable ~ s(raster variable, yday) + week" 14](#_Toc519356945)

# List of Tables

[Table 1 Formulas for modelling daily temperature at the different timeframes, where temp\_var is daily minimum, maximum and mean temperature. 3](#_Toc519356946)

# Introduction

Regional maps of climate and weather conditions are crucial in agricultural, forestry and viticultural research. In particular, the temperature of the air near the surface of the earth (~2m above surface, surface-air temperature, Ts) has a strong influence the growth and development of plants and insects. Ts maps are used in spatial analyses to select suitable cropland, estimate crop yield, predict pest population, and others that improve farming management practices. However, Ts data collection is limited to discrete points at weather station locations. Consequently, estimation in-between the stations (i.e. spatial interpolation) is necessary to produce a regional map of Ts maps.

A network of weather stations from Environment Canada (ECAN), the Department of Natural Resources (DNR), and the Applied Geomatics Research Group (AGRG) has been growing throughout Southwest Nova Scotia (SWNS) since the early 1900’s, reaching its peak in 2012. Data was available for this study at 98 stations (AGRG, 71; DNR, 11; ECAN, 16), which are shown in Figure 1‑1.

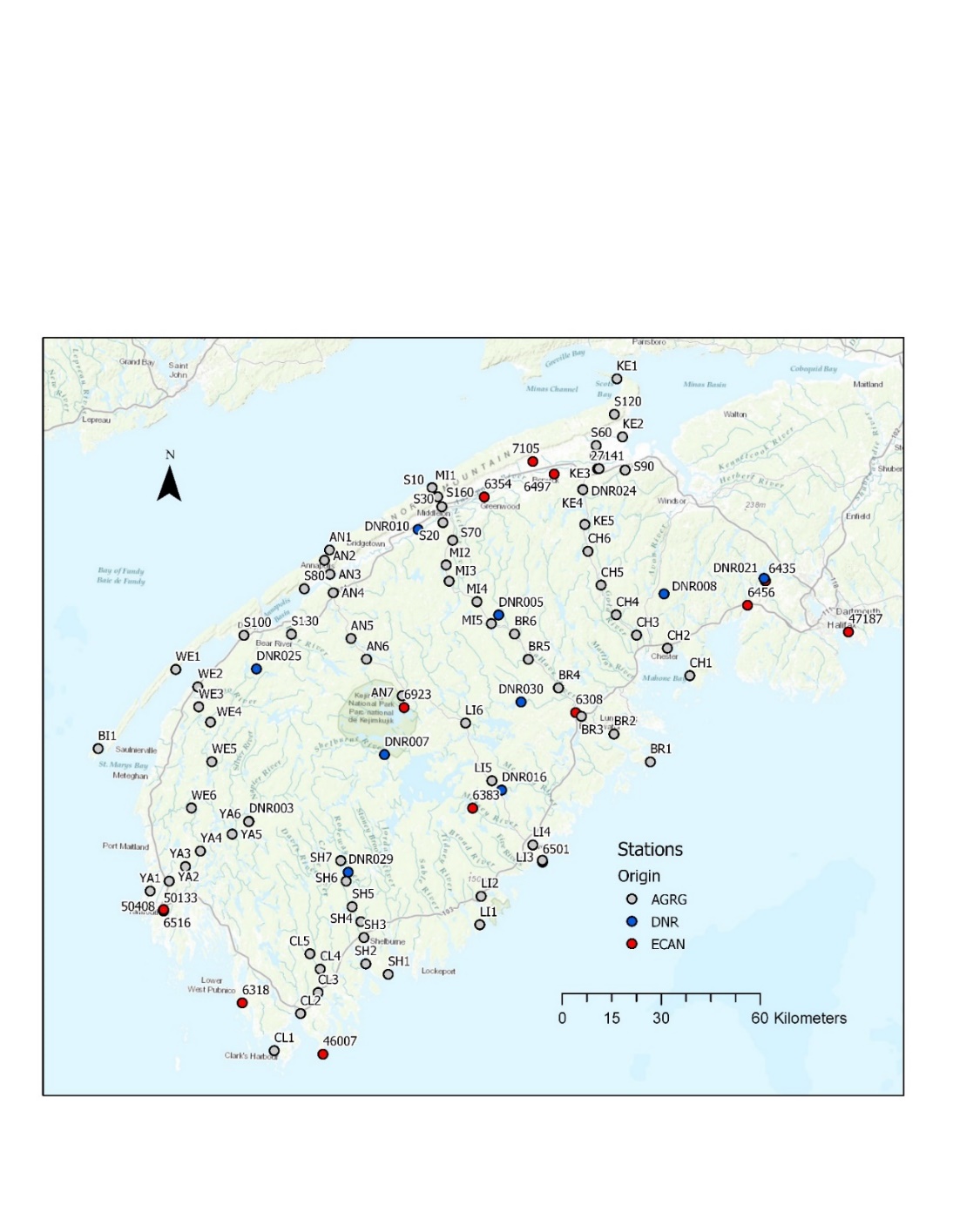
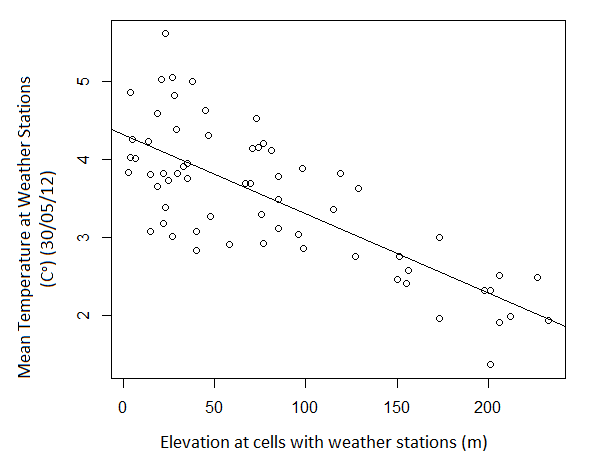
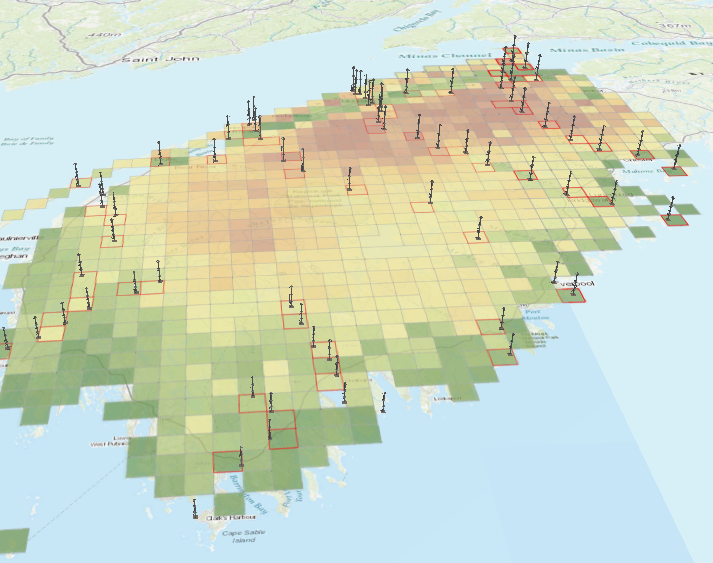


Figure 1‑1 Weather stations locations throughout SWNS. Origins are shown by colour: AGRG (grey), DNR (blue), and ECAN (red). Stations are labelled by their unique station I.D.

The Ts readings from the weather stations are stored as comma separate value files (.CSV’s), which as is, are difficult to extract information from. The objective of this project is to build upon the usefulness of the temperature readings by mapping the data throughout the entire region, by spatial interpolation. The goal is to generate maps of daily average Ts, and daily accumulating growing degree days (GDD), for the growing seasons (April – November) of the available complete years of sufficient data, 2012 – 2017.

Temperature maps were traditionally interpolated by hand-drawn isolines based on the researcher’s knowledge of the area. More complex and computationally heavy estimation procedures were developed and automated as technology progressed. Different procedures include simple mathematical models (ex: inverse distance weighting), regression techniques (ex: multivariate linear regression, splines, generalized additive models GAMs), and geostatistical methods (ex: Kriging).

GAMs are a non-parametric regression technique. With regression techniques, the relationship between dependent and independent variables can be evaluated in depth. A non-parametric model, like a GAM, does not rely on a predetermined function, and is therefore a completely data-driven, *a posteriori* technique. Natural phenomena, such as Ts, are unpredictable and highly variable, any assumptions made *a priori* may smooth over small nuances and detract from overall accuracy. An illustration of how raster values are extracted to weather station locations, and used to model temperature is shown in Figure 1‑2. The models are then used to back-calculate the temperature at all locations

b)

a)

a

a

Figure 1‑2: Illustration of how weather station data is modelled with rasters. a) Weather stations at corresponding cells of DEM. b) Temperature records plotting against DEM value with a linear regression line

In agriculture and related fields, Ts is often represented as a unit called growing degree days (GDD, Equation 1, below). This unit is the daily mean Ts, minus a threshold for development, then accumulated from a specified start date (often, the start of the growing season). The goal of this project is to generate daily GDD maps for SWNS, for the 2012 – 2017 growing seasons, as well as their six-year mean.

|  |  |  |
| --- | --- | --- |
|  |  | 1. |

Where n is the number of days to accumulate, is the average of the daily maximum and daily minimum Ts, is the temperature threshold for development, and is the length of days of accumulation.

Environmental rasters selected as candidates for modelling were: elevation, proximity to coast, solar radiation, aspect, slope, and topographic position index. These rasters were selected because they are related to the processes that heat or cool Ts. Elevation is negatively related to Ts. Air parcels expand as elevation increases, caused by the lowering of atmospheric pressure, and the release of energy causes Ts to decrease. The solar radiation received by a surface plays a major role in heating Ts. Actual observations of solar radiation are available from online archives of remote sensing data. The aspect and slope of surface also contribute to the amount of solar radiation received. In the northern hemisphere, southern facing slopes receive the longest duration of sunlight in a day. The slope of a surface is related to Ts as solar radiation is most concentrated when striking at a perpendicular angle. Topographic position index is a variable that incorporates the aspect, slope and elevation.

Easting and northing rasters are useful in Ts modelling to include the spatial component of Ts variability. The rasters in combination give the geographic coordinates at all locations in the study area. Coordinates are known at weather stations and can be incorporated into the model as an independent variable, then applied the entire area with the rasters.

The relationship between temperature and the input rasters may depend on the length of the modelling timeframe. Models will be generated with daily temperatures from differing lengths of time: daily, weekly, monthly, yearly, and all six-years. The daily minimum, maximum and mean temperatures will be modelled separately. Ultimately, daily mean temperature maps are desired to generate GDD maps. However, some studies in the literature model daily minimum and maximum separately, then average the map outputs for daily mean temperature. Both methods will be tested.

## Study Area

The study area is the southwestern flank of Nova Scotia, one of Canada’s Maritime provinces. The province is a peninsula and sits on the 45th parallel (Figure 1‑3). The center of Southwest Nova Scotia is approximately 44.3750° N, 65.0311° W).

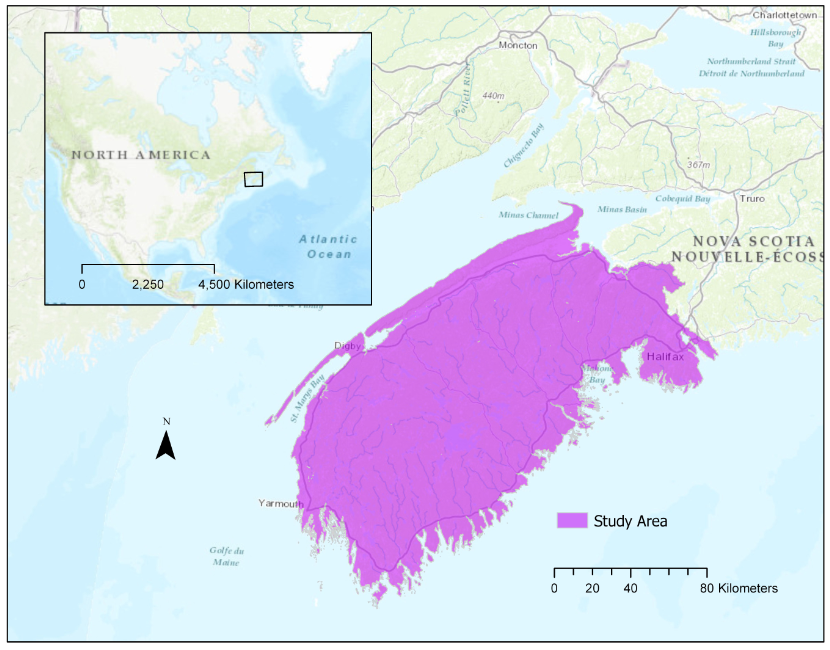


Figure 1‑3 Study area, Southwest Nova Scotia, Canada

The area is densely populated with coniferous and boreal forest, with areas of agriculture and vineyards. The area has several trees nurseries, apple orchards, vineyards and fruit farms (Government of Canada, 2017) (Figure 1‑4 &Figure 1‑5). There are approximately 3900 total farms in Nova Scotia, 39% of which are in the study area, SWNS (Statistics Canada, 2016). Kings County, in SWNS, has the most farms per county in Nova Scotia of oil, seed and grain; vegetables and melon; and fruits. The Lunenburg County, which is also in SWNS, has the most total nursery and tree production farms (Statistics Canada, 2016).

|  |  |
| --- | --- |
|  |  |
| Figure 1‑4 Relative fruit and tree nut farm in  Canadian Provinces 2016 (Stats Canada) | Figure 1‑5 Relative nursery and tree production  in Canadian Provinces 2016 (Stats Canada) |

Nova Scotia also has a large and growing wine industry. Within Nova Scotia, there are four main wine growing regions: The Annapolis Valley, Gaspereau Valley, South Shore and Magalash Peninsula. These regions, except for the Magalash Peninsula are located with SWNS (Figure 1‑6). Approximately $196 million to the Nova Scotia economy was brought in by the wine industry in 2011 (Frank, Rimerman + Co. LLP, 2013).

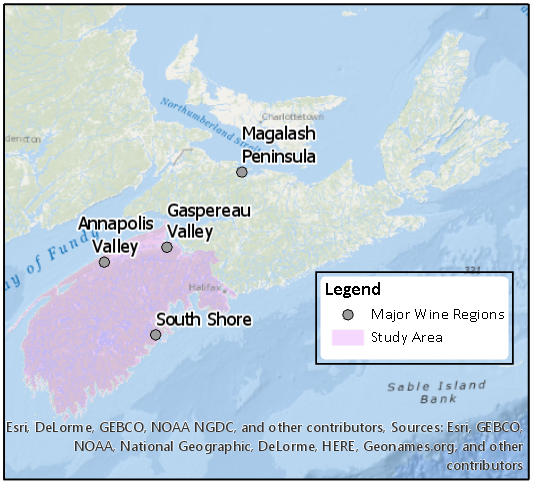


Figure 1‑6 The four major wine producing regions in Nova Scotia

# Methods and Data

## Data

#### Study Area ShapefileThe study area was defined by the outer boundary of the fourteen primary watersheds in SWNS (Figure 2‑1). The 1:10000 scale Nova Scotia watersheds shapefile was downloaded from the government of Nova Scotia’s open data portal (data.novascotia.ca). The polygons in SWNS were extracted from the shapefile in ArcGIS Pro, by manually selecting the polygons then using Copy Features to create a new subset shapefile.

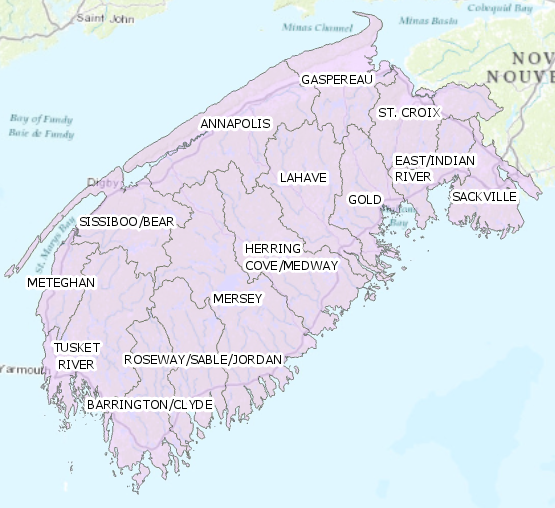


Figure 2‑1 The fourteen primary watersheds in SWNS that form the boundary of the study area.

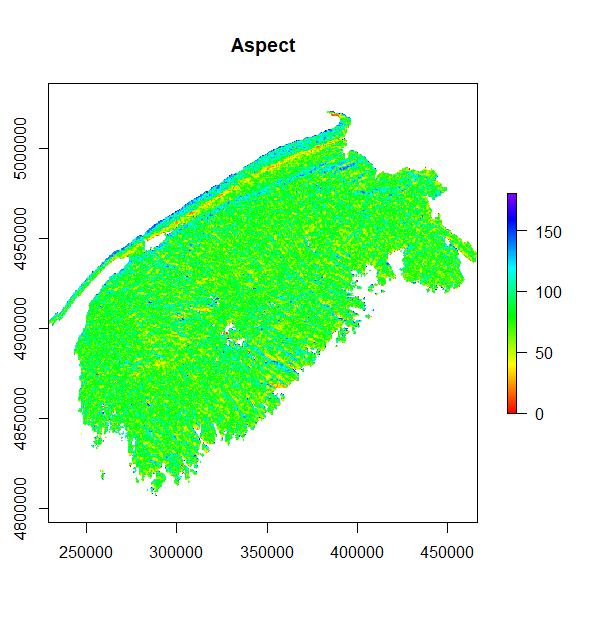
#### Stations Shapefile

The weather station location as a shapefile of points was acquired through the AGRG. Using the ArcGIS geoprocessing tool, *Add XY to Layer*, the geographic coordinates for each station in NAD 83 UTM Zone 20 were extracted to data table. These coordinates were joined back to the temperatures data frame by station I.D., to add geographic information to each station in the study.

#### Digital Elevation Model (DEM)

Figure 2‑2 The DEM of SWNS used in analysis. Reprojected, cropped and resampled from Nova Scotia Department of Natural Resources.

The enhanced DEM product (DP ME 55, Version 2, 2006) (Department of Natural Resources, Geoscience & Mines Branch) of Nova Scotia was obtained online through the provincial government website. The DEM was originally generated at a resolution of 20m with input coverages that include elevation points, hydrologic features and contour lines. The DEM was cropped to the study area in ArcGIS Pro with the *Extract by Mask* tool.



#### Aspect

Figure 2‑3 The aspect raster in used in analysis. Generated with the DEM in Figure 2‑2 and the Aspect tool in ArcGIS Pro

The aspect raster was generated from the SWSN DEM using the *Aspect* tool in ArcGIS Pro (Method = Planar). The following function was then applied to the raster cell values within R Studio (See Appendix A: Scripts for Data Preparation):

The aspect raster now has a range of 0-180, where 0 is completely northward, 180 is completely southward, and 90 is completely eastward or westward.

#### Proximity to the Coastline

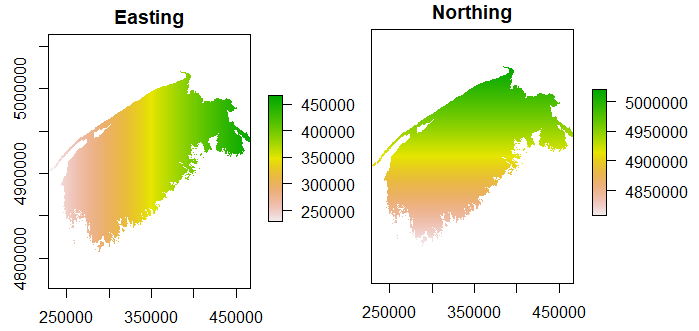
A generalized coastline was digitized around SWNS in ArcGIS Pro. Then the proximity to coastline raster was generated using the *Euclidean Distance* tool in ArcGIS Pro, and all environments set to the modified DEM (cell size, extents, coordinate system). Within R Studio, the values of the raster were capped at 30 000m (See Appendix A: Scripts for Data Preparation).

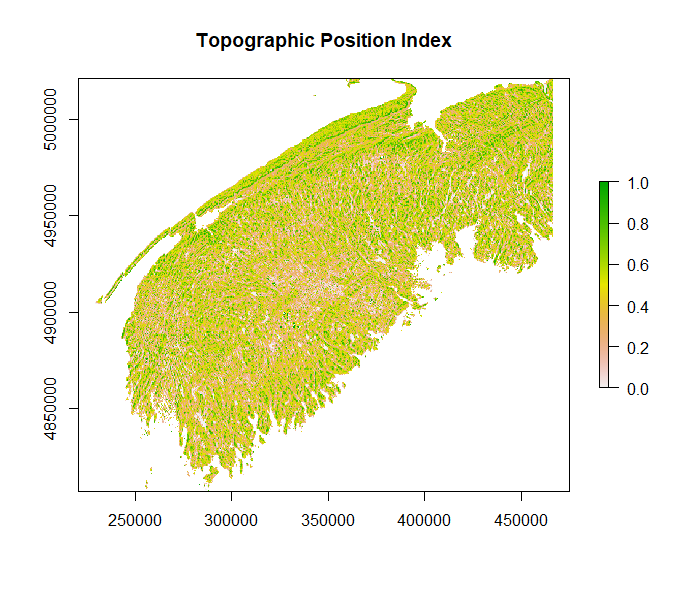
#### Slope

The slope raster was generated from the SWNS dem using the *Slope* tool in ArcGIS Pro (Output Measurement = Degrees, Method = planar, Z factor = 1).

#### Easting and Northing Rasters

An easting and a northing raster were generated from the SWNS DEM by a multi-step workflow in ArcGIS Pro. First, the DEM was converted to a points feature class with *Raster to Points*. Eastings and northing values were then extracted to the attribute table, with the *Add XY to Layer* tool. The points were finally converted back to a raster twice, once with the eastings then with the northings with *Convert Points to Raster*.





#### Topographic Position Index (TPI)

The TPI raster was acquired through the AGRG at a 20m resolution.

#### Solar Radiation

Solar radiation rasters were acquired as daily sums at a 700m resolution from the AGRG. The original source of the data is the GOES archive online (reference this better). Using arcpy functions, the rasters were reprojected, masked to the study area, and resampled to a 200m resolution (See Appendix A: Scripts for Data Preparation - 2 ).

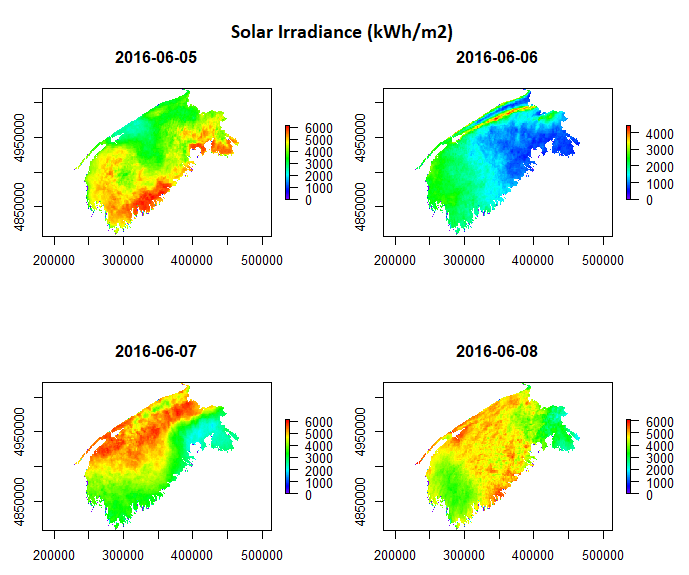


Figure 2‑4 The daily sums of solar irradiation (kW h/m2 ) in SWNS for the dates June 5th - June 8th, 2016.

## Method Stages Overview

The basic outline of the methodology is described in the flowchart in Figure 2‑5. In the first stage, input raster values were extracted based on weather station locations, and joined to the Ts data to create the modelling data frame. The suitability of each raster in the model was assessed by their variability and range at the weather stations, and the strength of their correlation in with Ts in individual GAMs. The weather stations were then divided into modelling stations, and validating stations. In the second stage, GAMs with different parameters were generated and tested. The parameters tested were: the dependent variables, the allowed number of knots (k), and the timeframe. The optimal model was chosen based on residuals at the validation stations, GCV scores, and adjusted R2 scores. In the third stage, the chosen GAM was applied to generate daily Ts maps. Accumulated GDD maps were then generated for each growing season (April – November) from 2012 – 2017. An R package, titled swnsmodelr, was written to facilitate various procedures in the methods.

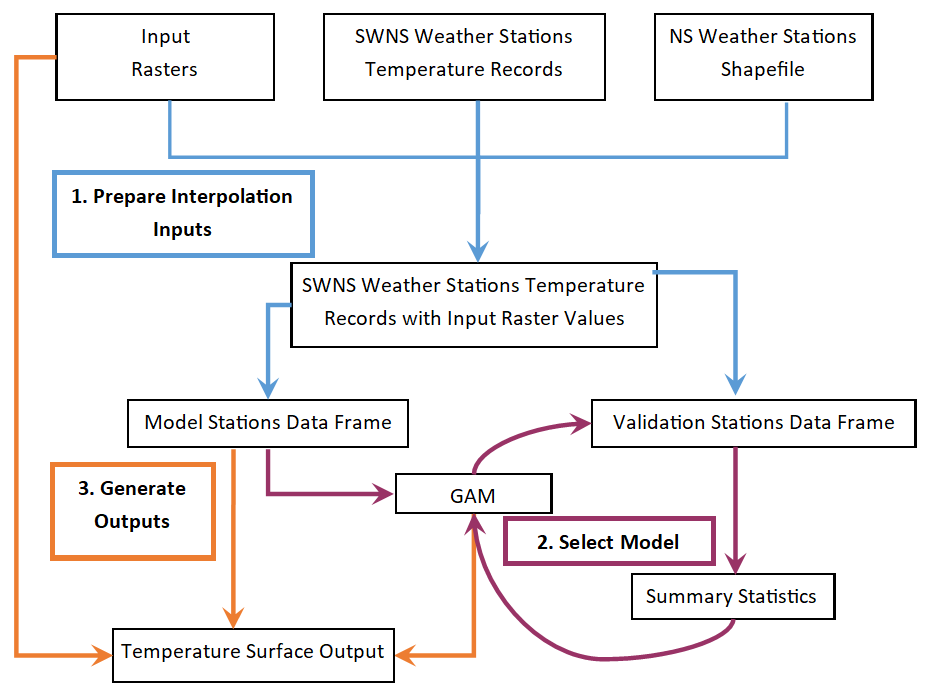
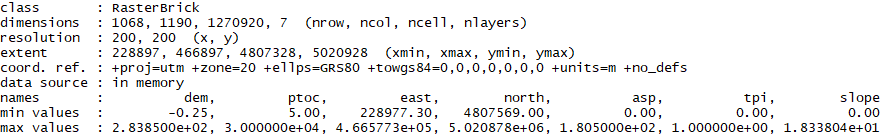


Figure 2‑5 Flowchart of the methods

## Stage 1: Prepare Interpolation Inputs

Preparing a data frame of the independent and dependent variables, and a *brick* of the input rasters are required for spatial interpolation in R. A raster brick is an R object of exactly congruent rasters (i.e. same extents, cell size, rows and columns).

### Raster Brick

The raster brick is an object with R that contains rasters, which on a cell by cell basis, the model will be applied to generate the interpolated output. The rasters must have cells that overlay exactly, i.e. the rasters must have the exact same cell size, number of rows and columns, and projected coordinate system, to have every cell aligning to calculate an output cell value (

).

Figure 2‑6 The structure of a raster brick object in R.

Notice multiple rasters with one set of dimensions, resolution, extent and coordinate system Although the input rasters do not need to be in a raster brick object until generating the final output, creating the raster brick first ensures that the rasters remain unchanged between the modelling and interpolation processes.

With a single function in R, raster::resample(), all the properties necessary to create a raster brick in a raster can be changed to match a second defined raster. The input rasters: DEM, easting, northing, aspect, slope and ptoc, were resampled with raster::resample() in R to a solar radiation raster (). The 2109 solar radiation rasters all have the same raster properties. The rasters are now able to form a raster brick object, generated with the function raster::brick(). The complete set of solar radiation rasters were not added to the brick at once, rather on a daily basis for each daily interpolated output.

.

### Data Frame

The temperature data was acquired as an Excel spreadsheet of daily weather stations records from 2012 – 2017 with the relevant columns: station I.D., date, mean temperature, maximum temperature and minimum temperature. The .CSV was loaded into R Studio as a data frame. The data frame requires the easting and northing for each weather station, and the corresponding values from each input raster.

The input rasters are separated into two categories for this process: constant rasters and temporal changing rasters. The distinction is made because the constant rasters (DEM, asp, slope, ptoc, easting and northing) only need to be extracted once to the weather station locations. Whereas the temporally changing rasters (solar irradiance rasters) need to be extracted to the weather station locations daily.

The easting and northing (metres, NAD83 CSRS UTM Zone 20N) of each weather station was joined to the temperature records data frame from the attribute table of the stations shapefile in R. The values of the constant rasters were extracted to the temperature records data frame with the function swnsmodelr::extract\_constant\_raster\_values(). The values of the solar radiation rasters were extracted to the temperature records with a combination of the functions swnsmodelr::make\_temporal\_rasters\_df and swnsmodelr::extract\_temporal\_raster\_values(). The final data frame has the daily temperature records for each of 98 stations, with the easting and northing values, as well extract values from the input rasters (Figure 2‑7).

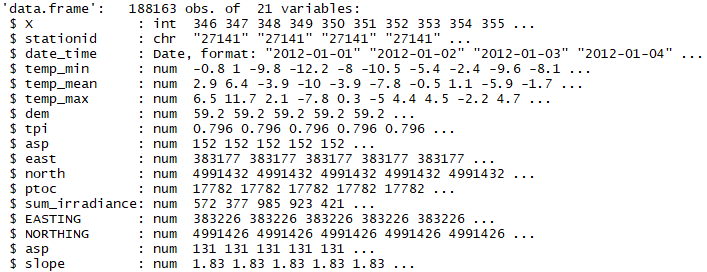


Figure 2‑7 Structure of final modelling data frame. Note: ‘east’ and ‘north’ are extracted from easting and northing rasters, whereas ‘EASTING’ and ‘NORTHING’ are actual coordinates at weather station.

### Separate Model and Validation Stations

The weather stations were assigned to one of two groups: model and validation. The data frame of temperatures and raster values at all weather stations, created in the previous section (Stage 1: Prepare Interpolation Inputs: Data Frame), was separated into two data frames based on the station assignments. The model station data frame was used to the generate the models, then the models were applied to the validation stations data frame to assess the accuracy.

Ideally, the AGRG weather stations would have formed the model stations, and the external stations (ECAN and DNR) formed the validations stations. The AGRG stations have been recording the mostly consistently since 2012, and have identical sensors among them. However, the AGRG stations lack placement in certain areas of Nova Scotia, therefore some external stations were included as modelling stations (See Figure 1‑1). Conversely, many external stations were very close to modelling stations, therefore some AGRG stations were included as validation stations (See Figure 1‑1).

The AGRG study stations alone do not provide sufficient coverage in the eastern most portion surrounding the Halifax area, and the Annapolis valley. There are external stations in those areas which were included in the model stations along with the study stations. The external stations that were included as model stations were ‘6354’ in Greenwood, ‘6456’ north is St. Margaret’s Bay, ‘47187’ in Halifax, ‘DN025’ south of Digby and ‘6497’ near Berwick. A few study stations were used in validating the models rather than generating them, these stations were: “YA4”, “S160”, “AN3”, “CL4”, “KE5”, “WE5”, “WE2”, “LI2”, “CH4”. These stations are in areas represented by model stations, but not so close that their modelling values were essentially the same.

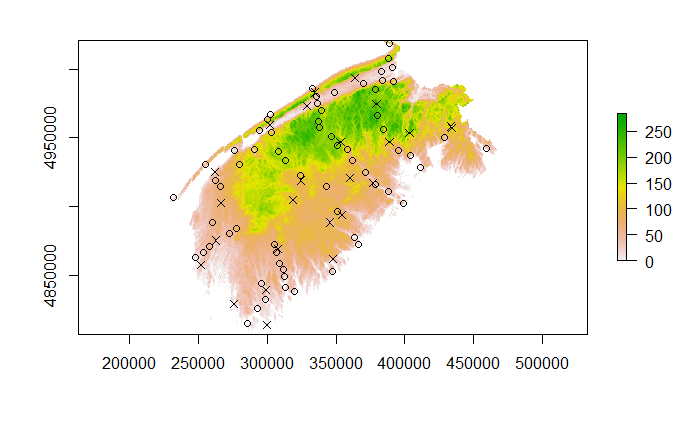


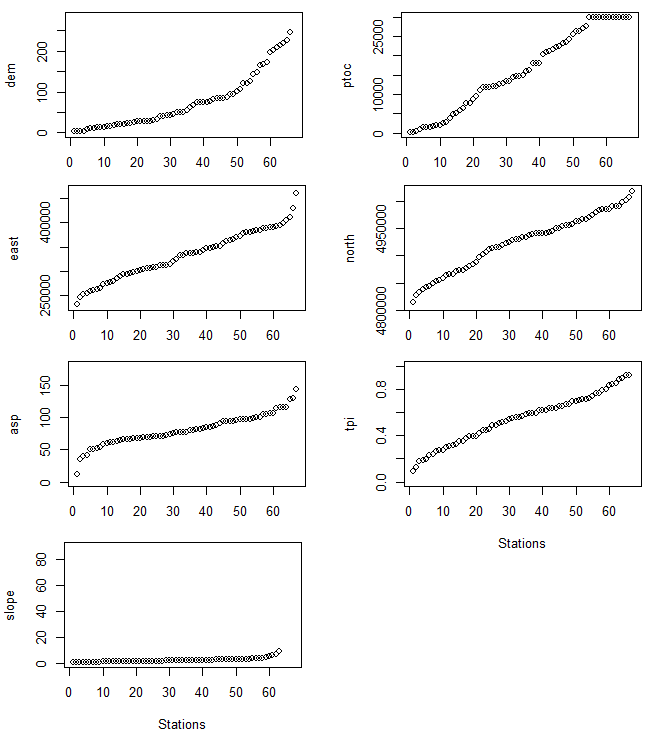
Figure 2‑8 The location of the model stations (circle) and validation stations (x) plotted over the SWNS DEM

## Stage 2: Model Selection

The suitability of considering each constant input raster for modelling (DEM, slope, aspect, easting, northing, tpi and PTOC) was assessed by plotting the range of raster values covered by the weather stations. Models were then tested by generating a series of test models and applying them to the 2012 temperature data. The parameters tested throughout for the selection process were: the daily temperature variable (mean, maximum or minimum), the allowed flexibility of the GAMs (i.e. allowed knots), and the modelling timeframe (all years, one year, monthly, weekly or daily).

### Raster Coverage by Weather Stations

The values of each constant input raster at the weather stations was plotted in order of raster value to assess the coverage (Figure 2‑9).

  
Figure 2‑9 The coverage of input raster values by weather stations.

Slope was under represented at the station locations, and was removed from the modelling process. The lack of slope values most likely stems from the stability of weather stations on flat surfaces rather than sloped surfaces. The other variables are relatively well covered, except for the high elevations (dem), the eastern most part of the study area (east) and the upper and lower ranges of the modified aspect. It is less likely that a surface is exactly north or south facing, so most of the aspect values are in the mid-range.

### Timeframe and Daily Temperature Variable

The daily Ts can be modelled with a single day of records from each weather station, or by using longer timeframes that can include multiple days of records. The modelling timeframes that were tested are: daily, weekly, monthly, annually, and the entire dataset (all six years).

There daily measures of Ts are available for each weather station: the daily maximum, minimum, and mean. The daily mean Ts was calculated from the maximum and minimum as in Equation [2]:

|  |  |  |
| --- | --- | --- |
|  |  | [2] |

The goal is to generate daily mean Ts rasters, which can either be done by direct interpolation of daily mean Ts, or by interpolating daily maximum and minimum Ts separately, then applying Equation [2] to the interpolated rasters. The latter requires twice as much interpolation, and will only be implemented if the results are significantly better than the former.

The different model formulas for the different timeframes that were tested separately on all three daily temperature variables are described in Table 2‑1.

Table 2‑1 Formulas for modelling daily temperature at the different timeframes, where temp\_var is daily minimum, maximum and mean temperature; yday is the day of year (i.e. 1 – 365).

|  |  |
| --- | --- |
| Timeframe | Model Formula |
| All six years | temp\_var ~  s(dem,month) + s(ptoc,month, k = 3) + s(sum\_irradiance, month) +  s(tpi,month) + s(asp, month) + s(east,north) + s(week, k= 3) +  year |
| One Year | temp\_var ~ s(dem,week) + s(ptoc,week, k = 3)+ s(sum\_irradiance, week)+  s(tpi,week)+ s(asp, week) + s(east,week) + s(yday, k=3) +  month |
| Monthly | temp\_var ~ s(dem,yday) + s(ptoc,yday k = 3)+ s(sum\_irradiance, yday) +  s(tpi,yday) + s(asp, yday) + s(east,north) + week |
| Weekly | temp\_var ~ s(dem,yday) + s(ptoc,yday k = 3) + s(sum\_irradiance, yday) +  s(tpi,yday) + s(asp, yday) + s(east,north) |
| Daily | temp\_var ~ s(dem) + s(ptoc, k = 3) + s(sum\_irradiance) + s(east,north) |

A raster term in parentheses with a date field (ex: s(dem, month) from the ‘All six years’ model) means that the values of the raster (i.e. the dem) are modelled against temperature while taking into consideration the date field (i.e. the individual months). A term without parentheses is a non-smooth term (ex: year from the ‘All six years’ model) by which the data is separated into pseudo sub-models within the model.

Each model was generated using data from the model stations, then applied to the validation station. The residuals from the models at the validation stations were collected, as well as the GCV scores, and adjusted R2 values from the models. For generating sample results for 2012, the all years model and annual model only need to be ran once. For the other timeframes (daily, weekly, monthly) functions were written for the swnsmodelr package to automate the model generation. The script used to

### Allowed Knots

Daily models can only have four raster variables due to restraints by the number of observations compared to the number of covariates. The p-values of modelling each raster variables separately with daily temperatures (Appendix A, Figure 0‑5) were used for deciding to exclude TPI and aspect.

## Stage 3: Generating Output

A daily mean temperature raster was generated for each day from 2012 – 2017. They were generated in R studio with the predict function of the raster R package. The function returns an estimated raster surface from a defined model formula and set of rasters that match the formula terms. Since the weekly models have day of the year in their formula, a raster for each day with all cell values equal to the day of year were generated and added to the set of rasters each day.

A GDD raster for each day was calculated from the daily mean temperature rasters. In R Studio, all daily mean temperature rasters were loaded into the environment as raster objects. The threshold temperature was subtracted from each raster, and negative values were set to 0. The rasters values were accumulated each day over each growing season. The procedure was repeated for thresholds of 5 and 10.

Six-year GDD means for growing seasons were calculated by averaging the accumulated GDD on the last day of the growing season, November 30th, for the six years.

# Results and Discussion

## Model Selection

### Timeframe

The summary statistics from modelling different timeframes suggest a divide between the two larger timeframes (all years and yearly) and the smaller timeframes (daily, weekly, and monthly). The residuals are least concentrated around zero for the two larger timeframes, followed by the monthly and weekly timeframes, then daily (Figure 3‑1).

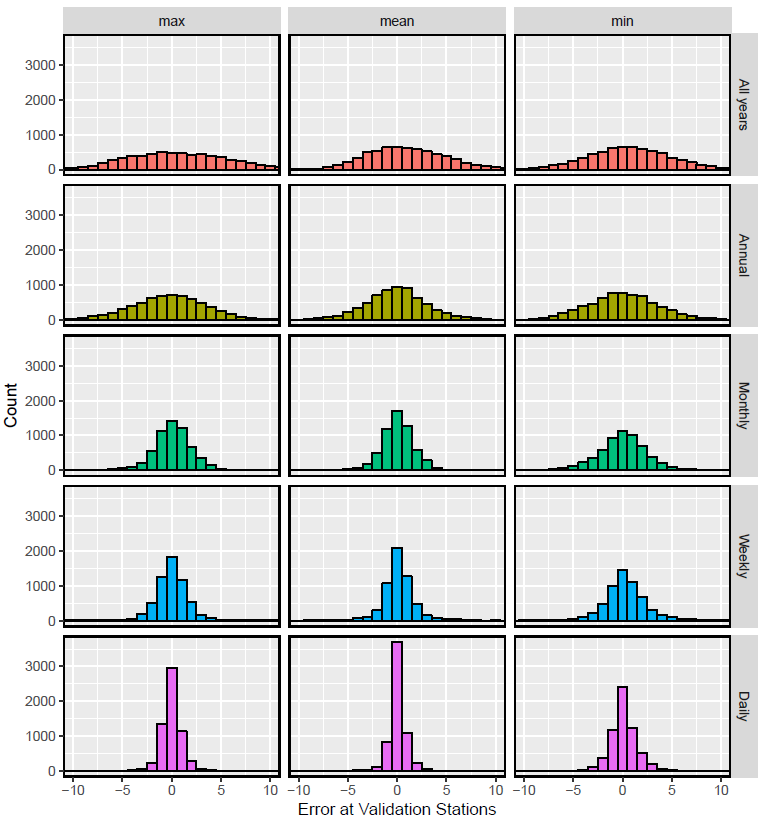


Figure 3‑1 Error at validation from models at five different timeframes

The distribution of residuals relative to daily temperature had a strong positive relationship in the larger timeframes, showing under- and over estimation of temperature lows and high, respectively (Appendix A, Figure 0‑2). This positive relationship was much subtler in the smaller timeframes.

The GCV scores reiterate that using the larger timeframes were potentially the least accurate. The larger timeframe GCV scores are much higher (approximately 15 – 20) than the smaller timeframes (< 4) (Appendix A, Figure 0‑3).

The adjusted R2 values for all years was and yearly were on mean 0.799 and 0.843, respectively. The weekly and monthly timeframes ranged from 0.724 – 0.963 and 0.611 – 0.984, respectively. They both followed a similar pattern throughout the year of starting and ending with high values and dipping in the middle of the year, however monthly models drop earlier (Appendix A, Figure 0‑4).

The combination of low adjusted R2 values and low GCV scores from the daily models suggests that they are more accurate despite having less significant correlation with the raster variables. Significance between temperature and raster variables is more likely to be found with larger dataframes where there are more records. It is possible that using larger timeframes is advantageous because of this feature, although in this case it may find false significance that is detrimental to the accuracy. The all years and yearly timeframes were not considered from here on the model selection process.

### Daily Temperature Variable

The models with daily mean temperature have residuals at the validation stations that are more concentrated around zero then daily minimum or maximum temperature (Figure 3‑1). The residuals at validation stations are distributed more randomly around zero in the daily mean temperature models; whereas the plots suggest under and overestimation of high and low values respectively in daily minimum and maximum temperature models (Appendix A, Figure 0‑2).

The daily mean temperature GCV scores lie closer to zero throughout the course of the year than the daily minimum and maximum models (Appendix A, Figure 0‑3).

These preliminary results do not suggest that there is advantage over modelling daily minimum and maximum temperature separately before calculating the daily mean. Therefore, only daily mean temperatures were considered from here on.

### Knots in smooth terms

Daily, weekly and monthly models of daily mean temperature were regenerated with k values of 1, 5, and 9. Counts of residuals at validations stations when k values are changed in smooth terms are shown in Figure 3‑2.

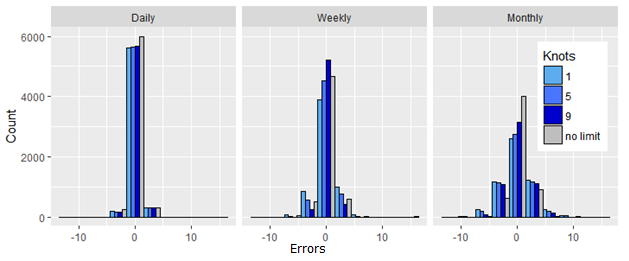
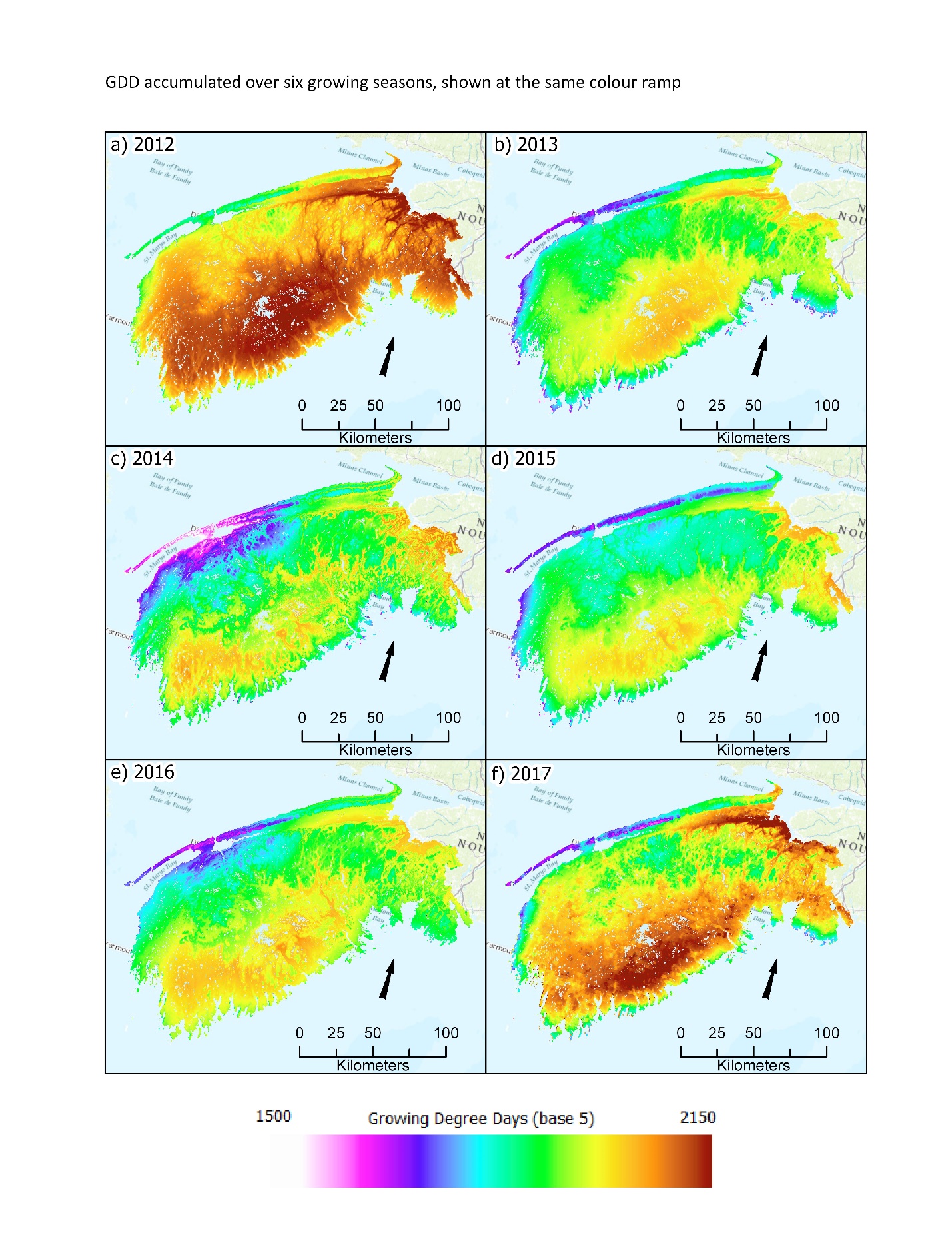
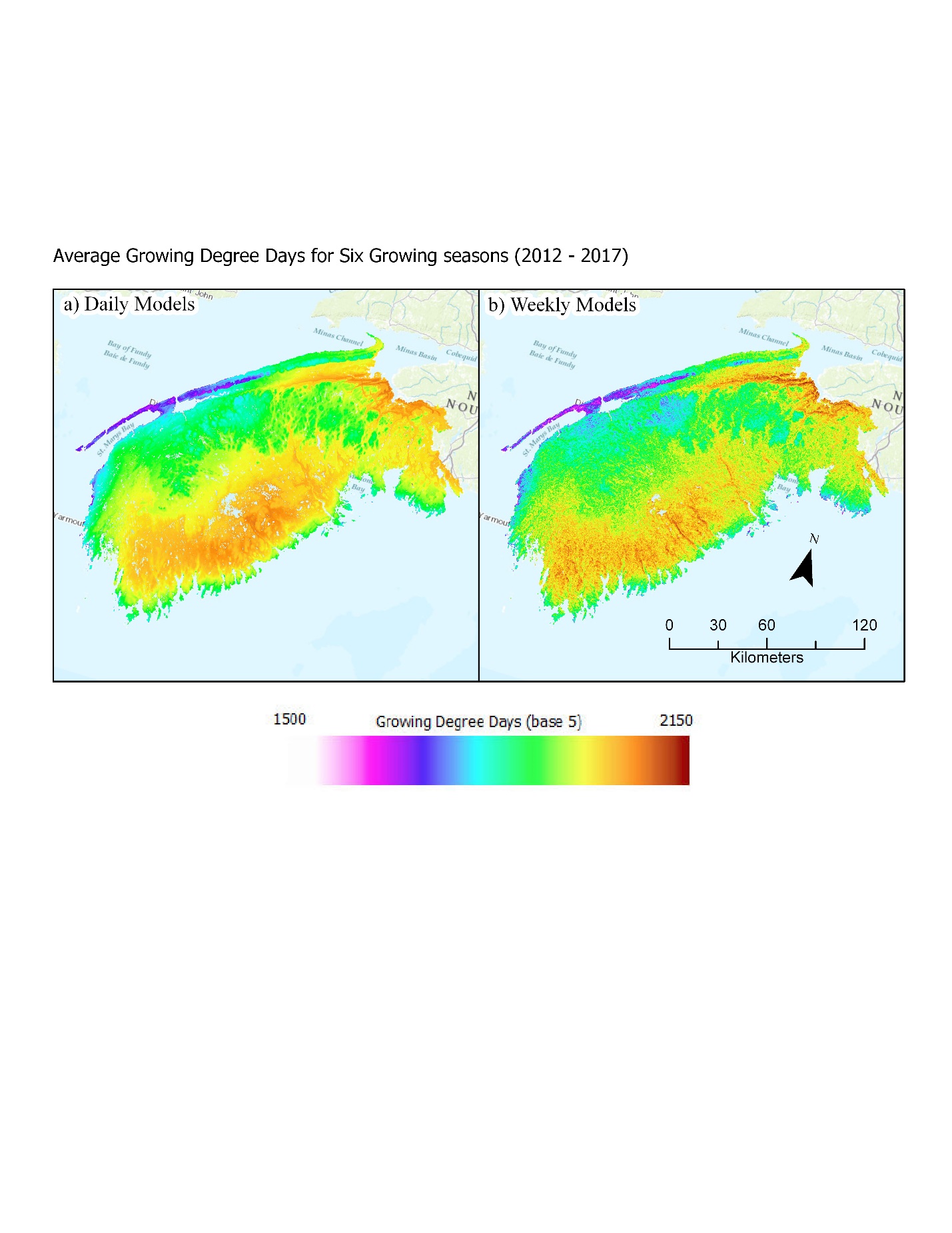


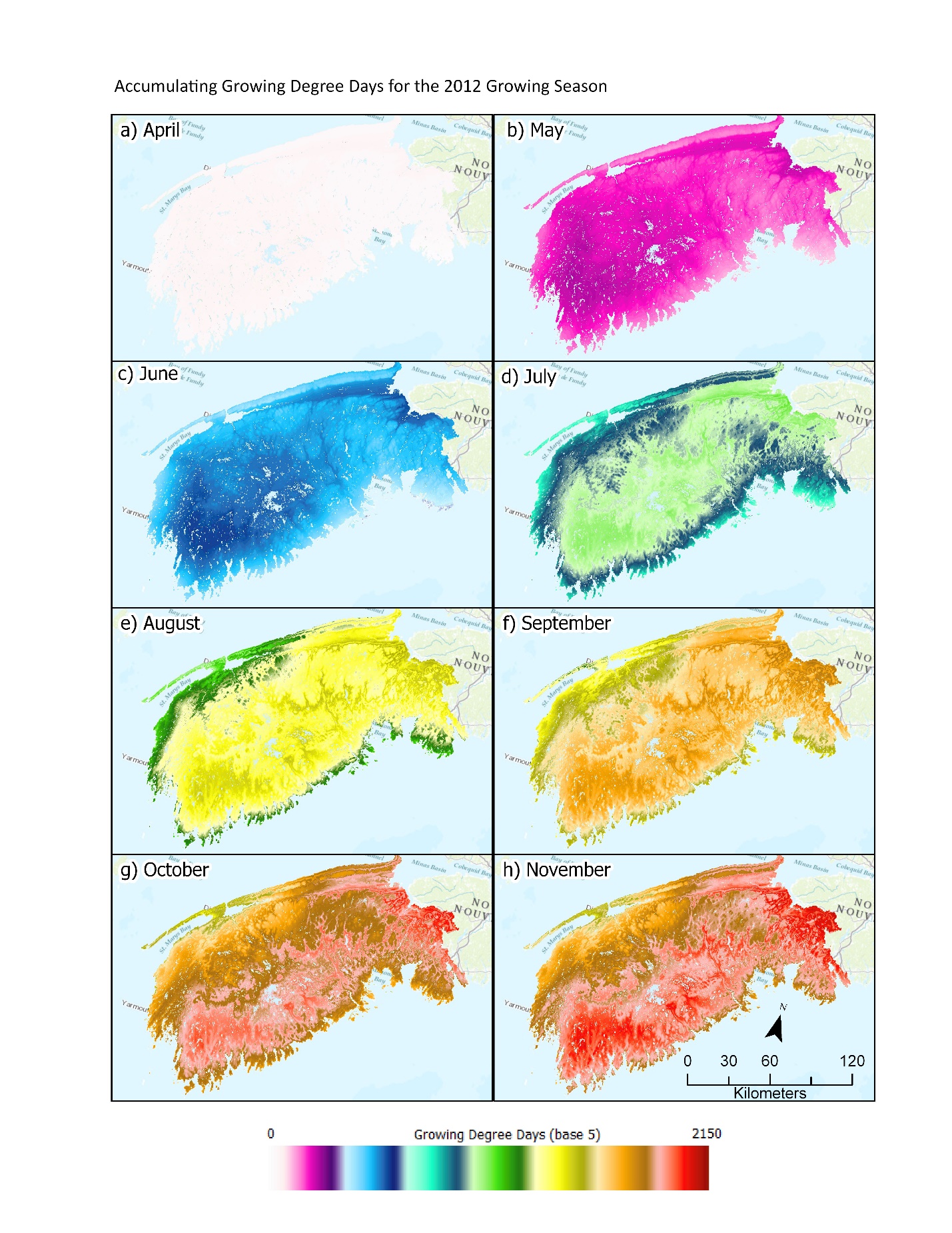
Figure 3‑2 Histogram of errors at validation stations from daily mean temperature models at three different timeframes (daily, weekly, month) with three different values of k in the smooth terms (1, 5, 9, no limit).

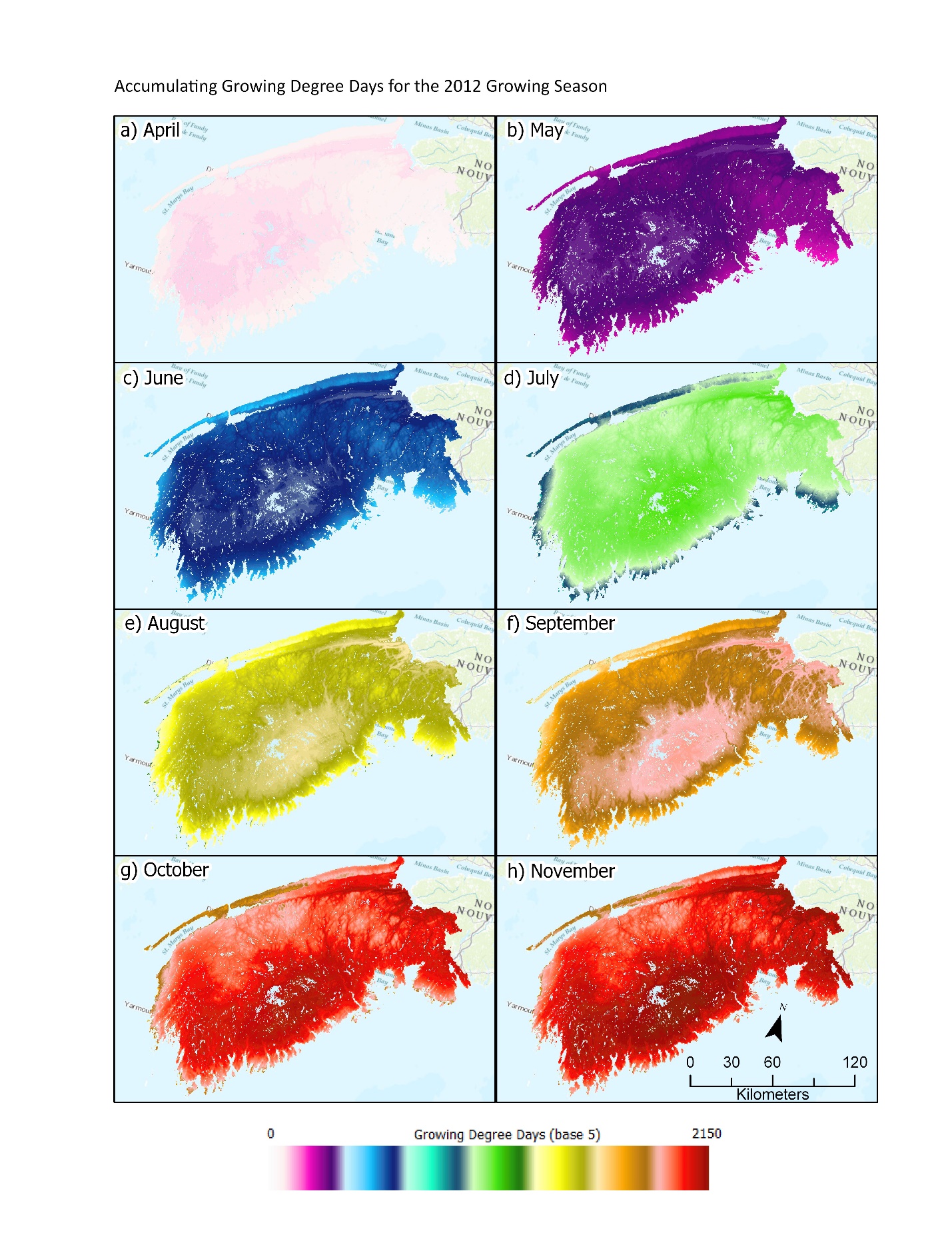
Limiting the knots in the daily timeframe slightly raised the residuals, with little difference between for the values of k. Residuals increased in weekly models when k was set to 1, stayed the same with k set to 5, and actually decreased with k set to 9. Setting knots in monthly models consistently increased residuals, with greater increases as knots were lowered.

## Accumulated Monthly GDD









# Appendix A: Sample Scripts for Data Preparation

These scripts were used throughout the methods to prepare the interpolation inputs.

## Modify Aspect and Proximity to the Coastline Rasters (R)

# Modify the Aspect and PTOC Rasters

library(raster)  
## Recalculate Aspect Raster   
asp\_raster\_in <- raster(file.path("Rasters","200","asp.tif"))  
asp\_raster\_out <- abs(asp\_raster - 180)  
writeRaster(asp\_raster\_out, file.path("Rasters","200","asp.tif"),  
 overwrite = TRUE)  
  
## Limit Proximity to the Coastline Raster  
limit <- 30000 # metres  
ptoc\_raster\_in <- raster(file.path("Rasters","200","ptoc.tif"))  
ptoc\_raster\_out <- ptoc\_raster\_in  
ptoc\_raster\_out[ptoc\_raster\_in >= limit] <- limit

writeRaster(ptoc\_raster\_out, file.path("Rasters","200","ptoc.tif"),  
 overwrite = TRUE)

## Resample Solar Radiation Rasters (Python)

# Resample all solar radiation rasters

import arcpy

arcpy.env.workspace = r"D:\GOES" # in folder

output\_workspace = r"E:\GOES\_200m" # out folder

rasters = arcpy.ListRasters() # create raster list

# Loop through list and resample to 200m

for i in range(0,(len(rasters)+1)):

raster\_name = rasters[i][:-4]

arcpy.management.Resample(r"%s.tif"%raster\_name,\

"%s\\%s.tif"%(output\_workspace,raster\_name),

"200 200", "BILINEAR")

print(raster\_name)

## Create Raster Brick

## Generate raster brick of constant rasters  
# List of raster names  
rasters\_names\_list <- list("dem", # elevation  
 "ptoc", # proxmity to coast  
 "east", # easting  
 "north",# northing  
 "asp", # aspect  
 "tpi", # topographic position index  
 "slope") # slope  
# List of raster objects  
rasters\_list <- lapply(FUN = raster,  
 X = paste0("E:\\Packages\\swnsmodelr\\Rasters\\200\\",  
 rasters\_names\_list,   
 ".tif"))  
  
# Brick of rasters  
rasters\_brick <- rasters\_list %>% brick()   
  
## Example of how solar radiation rasters are added based on date  
# Make dataframe of solar raster paths and dates  
solar\_rasters\_df <- make\_temporal\_raster\_df(file.path("E:","GOES\_200m"),  
 ymd("2016-01-01"),  
 ymd("2016-12-31"),  
 date\_chars = c(16, -5),  
 date\_format = "%Y\_%j")  
# Filter data frame for solar raster path based on date  
solar\_now\_df <- solar\_rasters\_df %>% filter(date\_time == ymd("2016-04-01"))  
# Generate raster from path  
solar\_raster\_now <- solar\_now\_df[[1]] %>% raster()  
  
## Add solar raster to brick  
rasters\_brick <- addLayer(rasters\_brick, solar\_raster\_now)  
  
## After modelling for date is finished, drop solar raster from brick  
rasters\_brick <- dropLayer(rasters\_brick, solar\_raster\_now)

## Prepare Data Frame

# Appendix C: Scripts Used in Model Testing

## Test multiple timeframes and daily temperature variables

# Testing models with different daily temperature variables and timeframes  
  
##### All years timeframe #####  
temp\_var <- list("min", "max", "mean")  
all\_years <- list()  
all\_years\_val <- list()  
for(i in seq\_along(temp\_var)){  
 all\_years[[i]] <- gam(formula(paste0("temp\_",temp\_var," ~  
 s(dem,month) +  
 s(ptoc,month, k= 3) +  
 s(sum\_irradiance, month) +  
 s(tpi,month) +  
 s(asp, month) +  
 s(east,north) +  
 s(week) +  
 year")),   
 data = model\_stations\_df)  
 all\_years\_val[[i]] <- add\_residuals(val\_df\_2012, all\_years[[i]])  
  
   
 if(is.na(all\_years\_val[[i]]$abs\_resid)){  
 all\_years[[i]] <- gam(formula(paste0("temp\_",temp\_var[[i]]," ~  
 s(dem,month) +  
 s(ptoc,month, k= 3) +  
 s(tpi,month)+  
 s(asp, month) +  
 s(east,north) +  
 s(week) +  
 year")),   
 data = model\_stations\_df)  
 all\_years\_val[[i]] <- add\_residuals(val\_df\_2012, all\_years[[i]])  
 }  
 all\_years\_val[[i]]$timeframe <- "All years"  
 all\_years\_val[[i]]$knots <- "No limit"  
 all\_years\_val[[i]]$temp\_var <- temp\_var[[i]]  
 all\_years\_val[[i]]$gcv <- all\_years[[i]]$gcv.ubre  
 all\_years\_val[[i]]$rsq <- summary(all\_years[[i]])[[10]]  
 all\_years\_val[[i]]$dev <- summary(all\_years[[i]])[[14]]  
 all\_years\_val[[i]]$abs\_resid <- abs(all\_years\_val[[i]]$resid)  
 for(l in seq\_along(summary(all\_years[[i]])[[7]])){  
 all\_years\_val[[i]]$var\_pval <- summary(all\_years[[i]])[[8]][[l]]  
 names(all\_years\_val[[i]])[names(all\_years\_val[[i]]) == "var\_pval"] <- names(summary(all\_years[[i]])[[7]])[[l]]  
 }  
}  
all\_years\_val$timeframe <- "All years"  
  
##### Annual timeframe ####  
annual <- list()  
annual\_val <- list()  
for(i in seq\_along(temp\_var)){   
 annual[[i]] <- gam(formula(paste0("temp\_",temp\_var[[i]]," ~  
 s(dem,week) +  
 s(ptoc,week, k= 3)+  
 s(sum\_irradiance, week) +  
 s(tpi,week)+  
 s(asp, week) +  
 s(east,week) +  
 s(yday) +  
 month")),   
 data = model\_df\_2012)  
 annual\_val[[i]] <- add\_residuals(val\_df\_2012, annual[[i]])  
   
 if(is.na(annual\_val[[i]]$abs\_resid)){  
 annual[[i]] <- gam(formula(paste0("temp\_",temp\_var[[i]]," ~  
 s(dem,week) +  
 s(ptoc,week, k= 3)+  
 s(tpi,week)+  
 s(asp, week) +  
 s(east,week) +  
 s(yday) +  
 month")),   
 data = model\_df\_2012)  
 annual\_val[[i]] <- add\_residuals(val\_df\_2012, annual[[i]])  
 }  
 for(l in seq\_along(summary(annual[[i]])[[7]])){  
 annual\_val[[i]]$var\_pval <- summary(annual[[i]])[[8]][[l]]  
 names(annual\_val[[i]])[names(annual\_val[[i]]) == "var\_pval"] <- names(summary(annual[[i]])[[7]])[[l]]  
 }  
 annual\_val[[i]]$timeframe <- "Annual"  
 annual\_val[[i]]$temp\_var <- temp\_var[[i]]  
 annual\_val[[i]]$gcv <- annual[[i]]$gcv.ubre  
 annual\_val[[i]]$rsq <- summary(annual[[i]])[[10]]  
 annual\_val[[i]]$knots <- "No limit"  
 annual\_val[[i]]$dev <- summary(annual[[i]])[[14]]  
 annual\_val[[i]]$abs\_resid <- abs(annual\_val[[i]]$resid)  
 }  
  
  
##### Monthly Timeframe #####  
# Create list of three dataframes  
monthly\_val\_list <- list()  
# Loop over each option: min, max and mean  
for(i in seq\_along(temp\_var)){  
 monthly\_val\_list[[i]] <- validate\_monthly\_GAMs(model\_stations\_df = model\_stations\_df,  
 val\_stations\_df = val\_stations\_df,  
 years = 2012,  
 months = 1:12,  
 formula =paste0("temp\_",temp\_var[[i]],"~  
 s(east,north) +  
 s(dem, yday) +  
 s(sum\_irradiance, yday) +  
 s(tpi, yday) +   
 s(asp, yday) +  
 s(ptoc, k = 3) +  
 week"),  
 alt\_formula =paste0("temp\_",temp\_var[[i]],"~  
 s(east,north, yday) +  
 s(dem, yday) +  
 s(tpi, yday) +   
 s(asp, yday)  
 s(ptoc, yday, k = 3) +  
 week")  
 )   
 monthly\_val\_list[[i]]$temp\_var <- temp\_var[[i]]  
}  
  
monthly\_val <- dplyr::bind\_rows(monthly\_val\_list)  
monthly\_val$timeframe <- "Monthly"  
  
  
##### Weekly Timeframe #####  
# Create list of three dataframes  
weekly\_val\_list <- list()  
# Loop over each option: min, max and mean  
for(i in seq\_along(temp\_var)){  
 weekly\_val\_list[[i]] <- validate\_weekly\_GAMs(model\_stations\_df = model\_stations\_df,  
 val\_stations\_df = val\_stations\_df,  
 years = 2012,  
 weeks = 1:52,  
 formula =paste0("temp\_",temp\_var[[i]],"~  
 s(east,north) +  
 s(dem, yday) +  
 s(sum\_irradiance, yday) +  
 s(asp, yday) +  
 s(tpi, yday) +   
 s(ptoc,yday)"),  
 alt\_formula =paste0("temp\_",temp\_var[[i]],"~  
 s(east,north) +  
 s(dem, yday) +  
 s(asp, yday) +  
 s(tpi, yday) +   
 s(ptoc,yday)")  
 )   
 weekly\_val\_list[[i]]$temp\_var <- temp\_var[[i]]  
   
}  
  
weekly\_val <- dplyr::bind\_rows(weekly\_val\_list)  
weekly\_val$timeframe <- "weekly"  
  
  
##### Daily Timeframe #####  
# Create list of three dataframes  
daily\_val\_list <- list()  
# Loop over each option: min, max and mean  
for(i in seq\_along(temp\_var)){  
 daily\_val\_list[[i]] <- validate\_daily\_GAMs(model\_stations\_df = model\_stations\_df,  
 val\_stations\_df = val\_stations\_df,  
 years = 2012,  
 days = 1:365,  
 formula =paste0("temp\_",temp\_var[[i]],"~  
 s(east,north) +  
 s(dem) +  
 s(sum\_irradiance) +  
 s(tpi) +   
 s(ptoc, k = 3)"),  
 alt\_formula =paste0("temp\_",temp\_var[[i]],"~  
 s(east,north) +  
 s(dem) +  
 s(tpi) +   
 s(ptoc, k = 3)")  
 )   
 daily\_val\_list[[i]]$temp\_var <- temp\_var[[i]]  
 print(temp\_var[[i]])  
   
}  
daily\_val <- dplyr::bind\_rows(daily\_val\_list)  
daily\_val$timeframe <- "Daily"

# Appendix B: Plots used in model selection

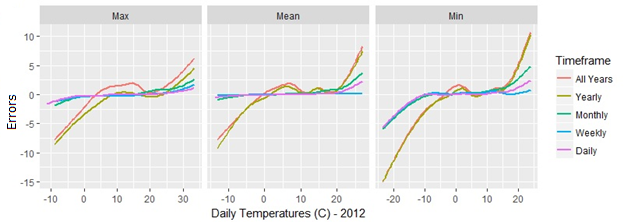


Figure 0‑1

Figure 0‑2 Distribution of residuals relative to three units of daily temperature (minimum, maxmimum and mean) at five different timeframes (all years, yearly, monthly, weekly, and daily)

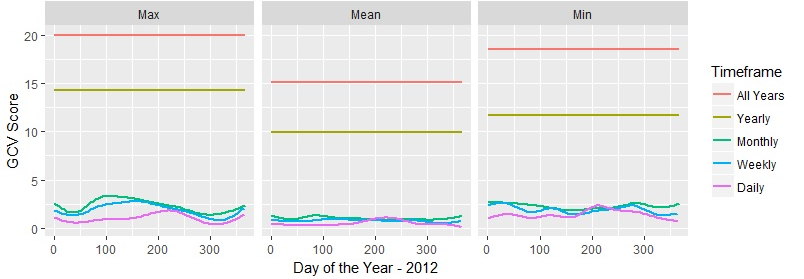


Figure 0‑3 GCV scores from modelling with three units of daily temperature (minimum, maximum and mean) at five different timeframes (all years, yearly, monthly, weekly, and daily)

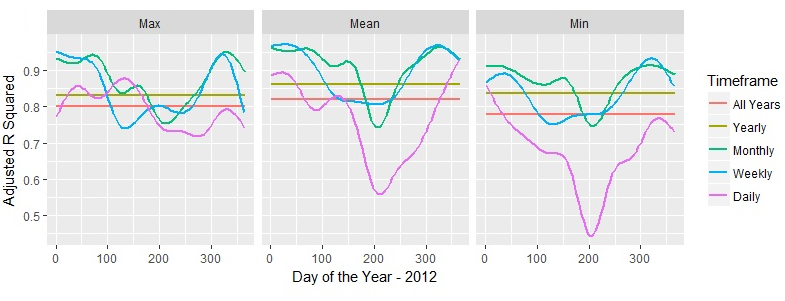


Figure 0‑4 Adjusted R2 values from modelling with three units of daily temperature (minimum, maximum and mean) at five different timeframes (all years, yearly, monthly, weekly, and daily)

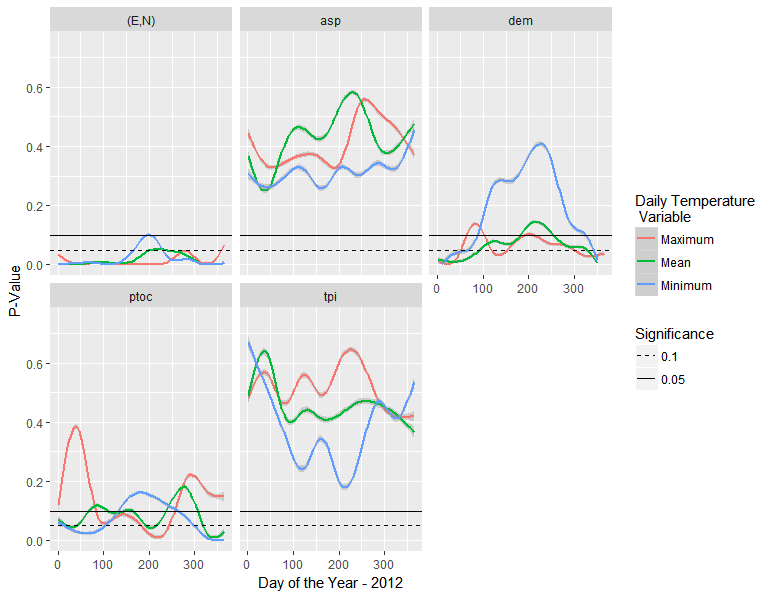


Figure 0‑5 P-Values of smooth terms from GAMs of daily temperature variables modelled daily against raster variables individually with the formula "temp variable ~ s(raster variable)"



Figure 0‑6 P-Values of smooth terms from GAMs of daily temperature variables modelled weekly against raster variables individually with the formula "temp variable ~ s(raster variable, yday)"

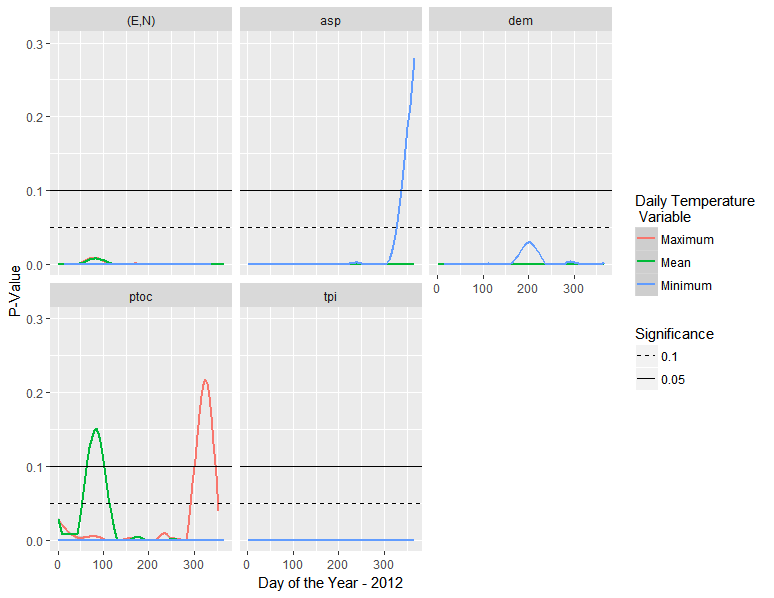


Figure 0‑7 P-Values of smooth terms from GAMs of daily temperature variables modelled monthly against raster variables individually with the formula "temp variable ~ s(raster variable, yday) + week"

When generating a GAM with mgcv, an important parameter to set is k for each smooth covariate. The value of k, which stands for ‘knots’, determines how close the GAM fits the data. At each knot the model can twist, meaning models with fewer knots are smoother, and more knots are noisy. Overall, the modelling parameters that need to be determined are: the daily temperature variable (minimum and maximum, or mean), the independent variables, the values of k, and the timeframe (daily, weekly, monthly, annual, all-years).

The statistics that will be used to judge the suitability of the parameters are the GCV scores, adjusted R2, and error at validation stations. The GCV scores are indicative of the residuals at the modelling stations. The adjusted R2 implies how well the environmental rasters as a set describe temperature variability in the model. As an empirical methodology, the error at validation stations ultimately gives the best sense of the accuracy of the models throughout SWNS.