GEOG309-22S2

Research for resilient environments and communities.

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Creating a contemporary climatological record of Cass Basin

Callum Davidson, Summer Jubb, Will Fraser, Xanthe Grainger, Aidan Nelson

Table of contents

Executive summary 3
1. Introduction 4
1.2 The study area 4
1.3 Importance of climate data. 5
1.4 Project stakeholders 5
1.5.1 Data sources 6
1.5.2 The Problem of Bogus Data 6
1.6 Non-linearity between stations 7
1.6.2 Valley pooling 9
1.6.3 Shading effect 9
..... 9
1.6.4 Adiabatic lapse rate 10
1.6.5 Synoptic setting 10
1.7 Selection of Imputation Scheme 10
2. Methods 11
2.1 Data formatting and storage 11
2.2 Data cleaning 11
2.3 Imputation Model: Rain 12
2.4 Imputation Model: Temperature 13
3. Results 13
3.2 Rainfall 14
3.5 Climate at Cass 17
4. Discussion 19
4.1 model success 19
4.2 Limitations 19
5. Conclusion 20
6. Acknowledgments 21
References 22
Appendix A 25
Appendix B 26
Appendix C 28

Executive summary

- This project poses the question “how can we use observational data, statistical techniques and research methods to improve the current climatological record at Cass Basin?”

- To this end, we have modelled a completed dataset of climate variables for the Cass basin to the present day to aid the Cass Management Research Area committee in research and ecological intervention in the basin.

- To do this the physical characteristics and climate of the basin were researched as well as the potential uses, stakeholders, and imputation methods for the dataset.

- Climate data from other stations in the vicinity of Cass was used to fill missing data in daily maximum and minimum temperature and weekly rainfall totals with a random forest machine learning imputation model.

- Overall, imputation proved to be relatively accurate, with performance largely dependent on the availability of data from nearby station.

- We hope that the methodological presented here might aid further research at Cass and may have applications in other places.

- The datasets, code, and report have been made publicly available via GitHub and drop box.

1. Introduction

Ground observations from weather stations remain the most reliable source of climatological data for any given area. The Cass Management Research Area committee wishes to have a complete climatological record for the Cass Basin area, so that recent trends and patterns can be used to inform projects such as the Cass native tree planting program as well as to aid future scientific research. There is limited research on how complex microclimates such as the Cass Basin are responding to a changing climate (Potter et al., 2013). However, the data sources that are currently available are filled with bad or missing data and only cover a few small periods in time. Understanding the long-term patterns in such an environment is essential to predict changes in physical processes and species distribution (Nowakowski et al., 2018). This project aims to generate a complete dataset containing an accurate representation of the historic climate at Cass. To this end, we conducted extensive research to identify potential stakeholders, topographical influences on the microclimate, and statistical methods for modelling and filling missing data. This report will outline this research and the methods we found has used to clean, format and fill climate data.

1.2 The study area

The Cass mountain research area (CMRA) is twelve square kilometres of land owned by the University of Canterbury (UC), situated in the wide valley and surrounding slopes of the Cass Basin. The Cass Basin forms part of the mid-Waimakariri intermontane river basin in the central South Island. The wide valley was formed by tectonic uplift, of which has been eroded by periods of glaciation and intense fluvial action (Perry et al., 1999). The eastern Southern

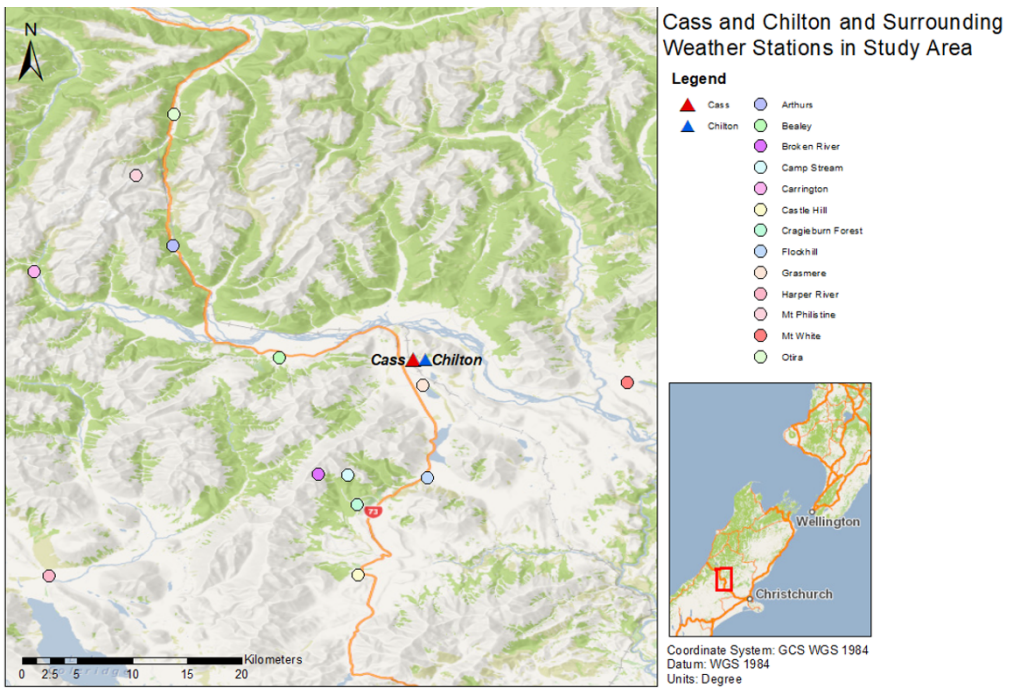

Figure 1: *Geographic reference to Cass Basin, including locations of the weather stations included in this study*

Alps orographic rain shadow has shaped the vegetation of the basin, historically Mountain and Black beech, which was destroyed by anthropogenic land clearing fires, increasing soil fertility for agricultural purposes (Perry et al., 1999). This induced the now predominant cover of short tussock grassland (Young et al., 2016). The dynamic environment is continuously active due to processes such as gravitational soil creep, tectonic processes (Burrows, 1977) and fires (Perry et al., 1999).

The Cass Basin has been home to at least two automatic weather stations (AWS): The Cass Station AWS is located at -43.03”S, 171.75”E at 572m a.s.l set up by UC in 1997, and the Chilton Valley AWS at -43.03”S, 171.76”E and 648m a.s.l ran from 1986-2005 (figure 1). There is a horizontal distance of 804m and elevation difference of 176m between the two stations.

1.3 Importance of climate data.

“Climate change” typically refers to the global mean surface temperature changes over time (Wake, 2015). Although global trends play an important role, local and regional microclimate processes can drive additional variability at small scales. Complete long term records of climate data from complex environments can help understand how these unique environments respond to global climate change. Academic literature has highlighted the importance of climate data as a foundation for many decisions, analyses, and tools (Pohl et al., 2022). The accuracy and accessibility of climate data is crucial due to the increasing number of people accessing the data and it being used for “billion-dollar decisions” (Overpeck et al., 2011).

1.4 Project stakeholders

Other than our primary stakeholder, the CMRA committee, the groups that have been identified as having interest in the project's findings include mana whenua, future research groups, farmers, recreationalists, and fire and emergency. The CMRA committee initiated this project as part of their wider goal to investigate the long-term variation in the microclimate Cass and are expected to utilise and expand this dataset for future research. The results have potential to contribute to projects such as the Cass native tree planting program run by Associate Professor David Evison.

An important aspect in conducting research in New Zealand is engaging with Māori; the original kaitiaki or custodians of the land (Rauika Māngai, 2020). The *guidelines for engagement with Māori* (N.d) state that one can build a relationship of trust and confidence by establishing contact early in the process of engagement whilst keeping an open mind. Following this advice, we met with Dr Abby Suszko to investigate a relationship between CMRA and Māori, we learnt that the relationship is in its infancy, therefore we will engage through a UC contact. *He Rautaki mō te Huringa Āhuarangi climate change strategy* (2018)

states that understanding upcoming changes is essential when making good decisions for future proofing the environments that provide for Ngāi Tahu's rangatahi. Temperature and rainfall readings are of particular interest, relevant to the climate change strategy, with food security and natural hazards being of primary concern regarding hauora.

1.5.1 Data sources

There are several weather stations that have been operated or are operating near to Cass. Few of these stations have a complete long-term record of data, with many gaps due to sensor malfunction, lack of maintenance, or other issues. Weather station datasets were collected from a range of sources including, the community partners' private collection at the university (for Cass and Chilton), NIWA's national database (Cliflo), and from Environment Canterbury's archives (by request). Stations recording precipitation and temperature are shown below in figure 2a and 2b respectively.

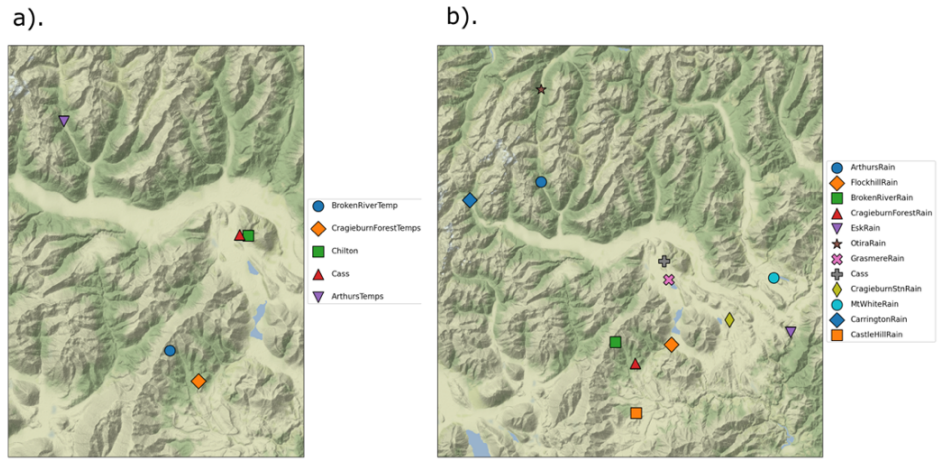

Figure 2: *a). Locations at which the temperature weather stations are located and b). Map of rain weather stations.*

We established a connection with Eva Nielson from UC, who has suggested that we could use small resolution European Centre for Medium-Range Weather Forecasts (ECMWF) reanalysis data as a data source, or to validate results. While this option has not been explored in this paper due to time constraints, assimilation of ERA5 data into a model might be useful in a future iteration of this project.

1.5.2 The Problem of Bogus Data

The accuracy of climate data is important to establish accurate trends in climate processes. Unfortunately, both Cass and Chilton are full of spurious data due to sensor malfunction and other issues. Using this data without careful examination can lead to completely inaccurate research conclusions, for instance Figure 3a) shows the polar temperature and wind plot for 1993 at Chilton station, compared to figure 3b) which displays the more accurate representation of the climate using the 1995 dataset. An unknown issue with the data

recorded between 1991-3 meant that temperature values below 0 were recorded as positive instead of negative, thus displaying the cold southerly winds as positive temperatures. The removal of such data will be essential to ensure the accuracy of our completed results.

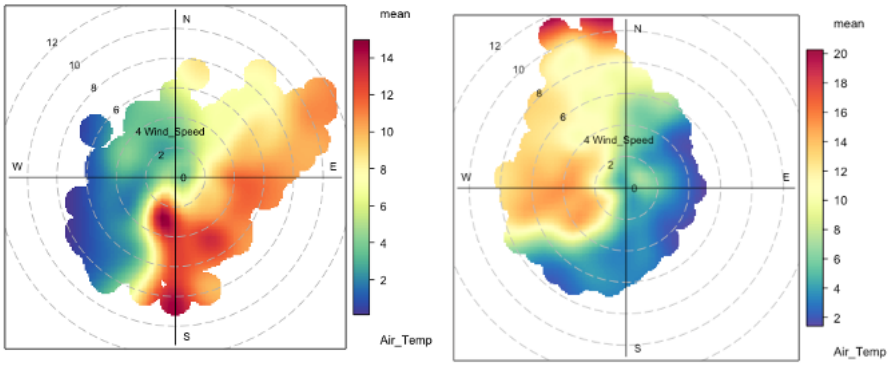

Figure 3: *a). polar plot for the uncleaned Chilton 1992 data and b). polar plot for the Chilton 1995 dataset.*

1.6 Non-linearity between stations

A wide range of literature has highlighted that topographic features and vegetation cover drive microclimatic variations and unique local atmospheric processes of an area (Qing-Ling et al., 2015). These microclimate processes can vary significantly across even small distances in Arthurs Pass.

The microclimates of New Zealand's alpine environments are a complex combination of boundary layer stability and energy flow characteristics (Sturman et al., 1985). Radiative energy exchange (REE) between the atmosphere and earth's surface drives local climate processes. The Surface Radiation Balance (SRB) refers to the incoming shortwave and outgoing longwave radiation fluxes at the surface (Suttles et al., 1986). Variations in the SRB are a function of the surface albedo, aspect, and inclination of the local slopes and interference from surrounding topography (Whiteman et al., 1989). Slope and valley winds are generated when the earth's surface heats and cools on an angle generating horizontal pressure gradients (Whiteman, 1990). These generate a range of localised weather processes such as cold air valley pooling, slope and valley winds (Whiteman et al., 1993), resulting in a highly variable and complex climate.

1.6.1 Wind channelling

The dominant wind direction and speed at Cass and Chilton are significantly different, even though they are situated only 800m apart (figure 4). The surrounding topography shelters the Chilton AWS significantly reducing wind speeds. However, Cass Station is exposed to the

predominant wind funnelling down the valley from the northwest, experiencing much faster wind speeds of up to 30ms^{-1} compared to the peaks of 6ms^{-1} of Chilton station.

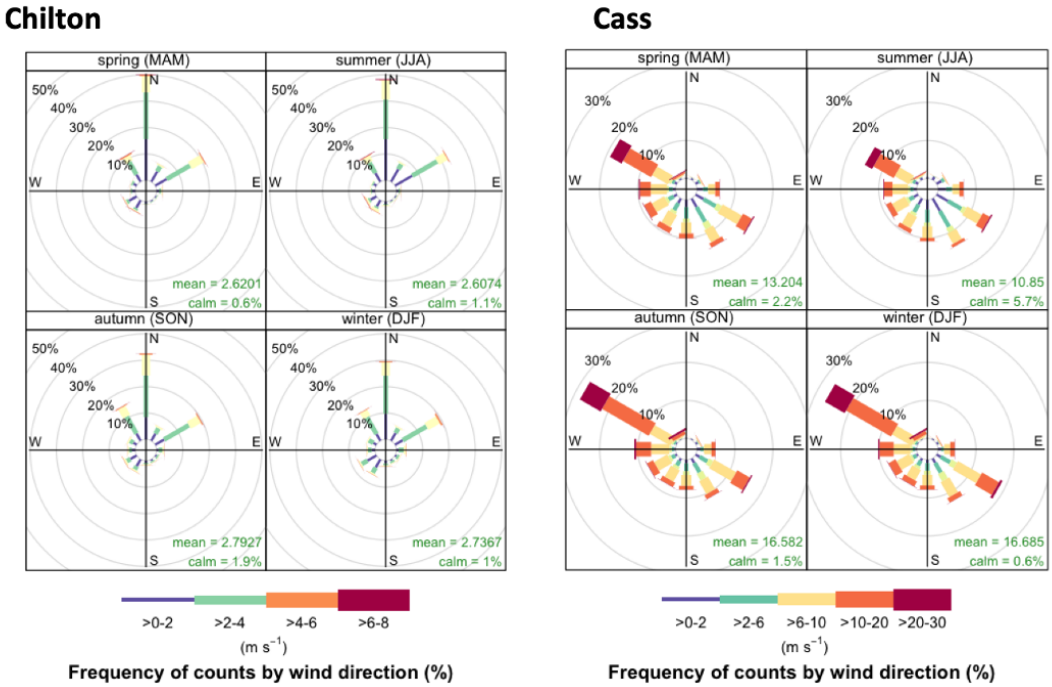

Figure 4: Windrose displaying the wind speed and direction at Cass and Chilton AWS. Figure generated in R studio.

The prominent wind flowing through the valley is warm from the northwest and cool from the southeast (figure 5). This reflects the typical synoptic weather patterns of the southern hemisphere's midlatitudes.

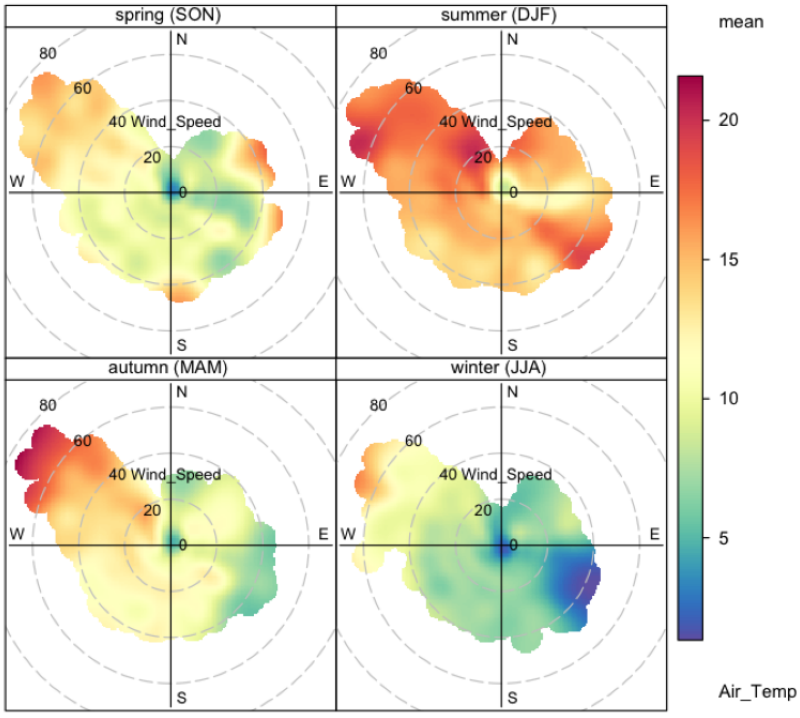

Figure 5: Polar plot displaying the wind speed, temperature and direction at Cass Station from 1997-2022.

1.6.2 Valley pooling

Valley pooling occurs within the Cass Basin due to stable temperature inversions forming as cool air sinks to the valley floor overnight (Lundquist et al., 2008). This can be seen be the cool downslope slow northerly wintertime winds seen in Figure 5. The effect of this is that the cold air pooling makes temperature readings on the valley floor significantly lower than stations at slightly high altitudes, above the inversion that has been formed. This process can generate complex non-linear relationships between temperature readings at nearby stations located at different heights, such as Cass and Chilton.

1.6.3 Shading effect

The mountainous topography northeast of Cass Basin influences incoming solar radiation at Cass and Chilton uniquely. The low angle of the sun in conjunction with the peaks northeast influences differential shading, whereby Cass station remains in the sun longer than Chilton (figure 6 and 7).

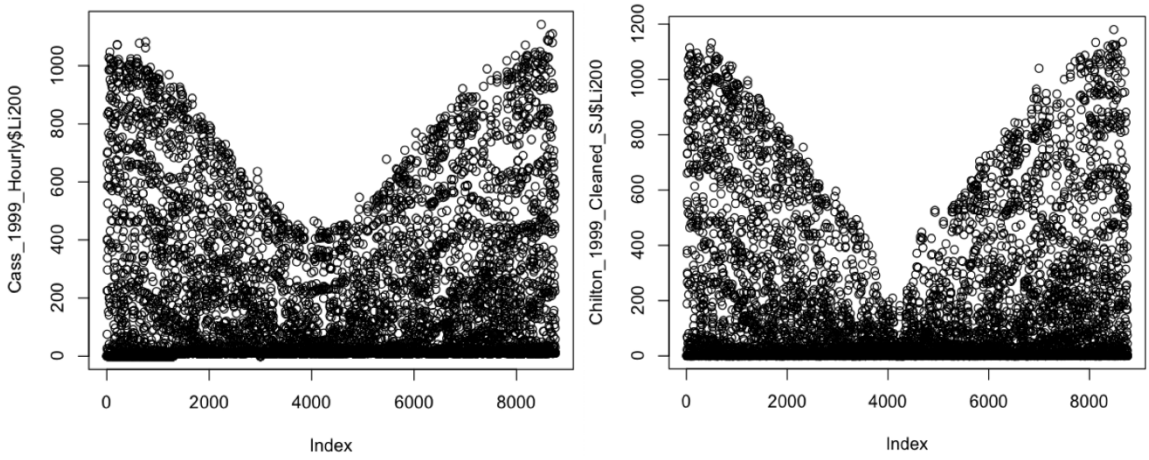

Figure 6: *General solar radiation at a broad-spectrum wavelength (Li200) at Cass and Chilton stations.*

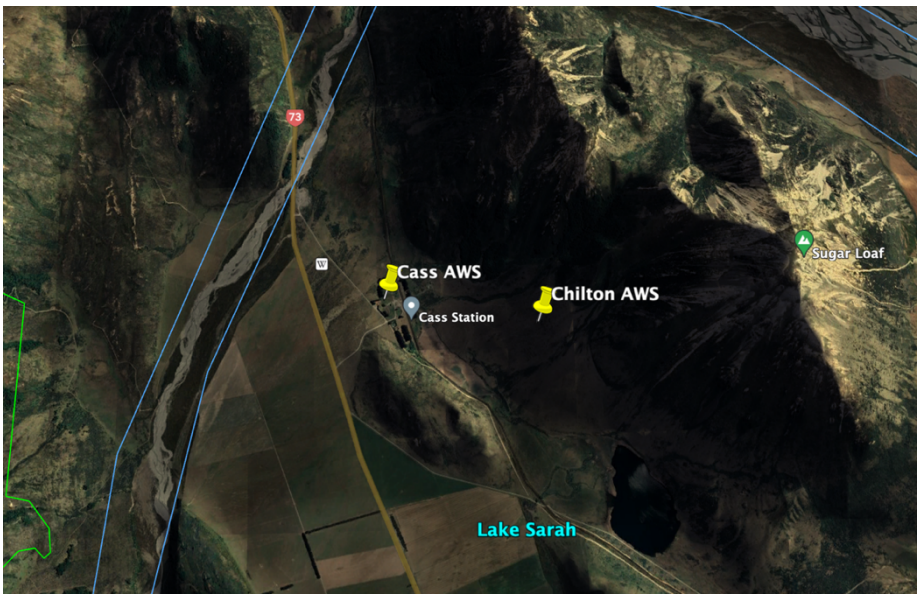

Figure 7: *Topographical shading at Cass and Chilton weather stations in the winter. Source: Google Earth*

1.6.4 Adiabatic lapse rate

The maximum and minimum temperatures at each station would be expected to follow similar fluctuations of seasonal temperature, only at different magnitudes. The adiabatic lapse rate of 9.8^{o}C per 1000m decreases the temperature at weather stations situated in higher elevations, as they are subject to more cooling with altitude (Blandford et al., 2008).

1.6.5 Synoptic setting

Thermally generated local climate features such as slope, valley winds and cold air pooling only occur in the Cass Basin when stronger synoptic scale winds are not present (Greenland, 1977). The Southern Alps synoptic setting is characterised by the interaction between the polar westerlies to the south, and the subtropical zone in the north (Greenland, 1977). This gives rise to a cycle of high-pressure systems and associated calm conditions being displaced by low pressure westerly flow on an average of every six to ten days (Soons, 1968). The predominant wind at Cass Basin is from the northwest, (McGowan & Sturman, 1996), occurring when the humid air of the Tasman Sea overtops the Alps dropping it's moisture content leaving warmer, drier airflow in the lee (Elvidge & Renfrew, 2016; Soons, 1968).

1.7 Selection of Imputation Scheme

Imputation is the filling of gaps using a statistical model to estimate missing values. There is a hierarchy of methods used in data imputation, ranging from very simple methods like single-value replacement of missing values, to very complex methods involving machine learning techniques. Given the suggested use-cases of our data, a model-based method will be necessary to ensure the accuracy of the datasets. There are many potential regression models that can be used, however the non-linearity between weather stations and high rates of missingness mean that some methods, such as linear regressions, will prove substandard. This is because linear regression assumes homoscedasticity (Tranmer & Elliot, 2008), of which the error of rainfall changes significantly across independent values.

Our data also has a high degree of missingness and very little or no overlap between some stations. There are also likely to be some amounts of outliers due to sensor failures. This is particularly true for rain gauge measurements, which are prone to blockages by debris or even becoming homes for arachnids, as was noted by the staff responsible for Chilton in 2002:

> *There is no rainfall recorded since 6 Feb as there was a spider web in the rain gauge.*

The missingness of the datasets can be overcome by using an iterative imputation process wherein data is imputed in a round-robin fashion, with imputed missing values used to build and improve the models used by the next round of imputation. This process encompasses a broad range of techniques that fall under the umbrella of Multiple Imputation. There are

several methods highlighted by academic literature, but two papers provide important insight to our project. Aguilera and colleagues (2020) compare various imputation techniques on rain gauge data with a high rate of missingness, sometimes in excess of 90%. They note the success of two techniques, spatio-temporal kriging and an implementation of Brieman's (2001) Random Forest regressors in a sequential imputation scheme. A second paper by Wolfensberger (et al., 2021) corroborates the success of The Random Forest algorithm for precipitation data over Switzerland, a similarly mountainous area like Arthurs Pass. Random Forest imputation has also been successfully employed in the imputation of temperature data by Zhang et al. (2022).

Random Forests (hereafter RF) is an ensemble learning methodology where a number of weak learners are fit to bootstrapped samples of the dataset who then vote on the best estimate of a missing variable. In the case of RF, the ensemble of weak learners is a collection of binary decision trees trained on different samples of the dataset. Missing values produced from one round of training and voting are used to train the next ensemble of learners in the successive iteration. This process is discussed in detail by Stekhoven & Bühlmann (2012) who successfully employed it in their MissForest algorithm used in this paper. The principal advantage of RF is that it is inherently non-parametric, meaning no assumptions about distributions or correlations between the datasets are made. It can also impute categorical datasets as well and it has relatively straightforward to tune hyperparameters, unlike some other machine learning techniques. Preliminary testing showed that it outperformed many other common imputation techniques including KNN neighbours (Mucherino et al., 2009) and parametric multiple linear regression (Azur et al., 2011).

2. Methods

2.1 Data formatting and storage

Due to the myriad of data sources and standards, the raw datasets needed reformatting to make them easily usable for analysis. Furthermore, previous stewards of the Cass and Chilton datasets have used inconsistent file formatting and file storage solutions. Variable names would change from one year to the next, and datasets were stored in everything from Microsoft Access 97 .mdb files, to Excel .xls spreadsheets and .csv text file with no column headings. The datasets were first reformatted to an internally consistent standard, with identical variable names (see appendix C). They were then all saved as comma delimited .csv files to ensure that they are easily accessible by a wide range of software and are robust to future changes in technology.

2.2 Data cleaning

A masking process was applied to remove physically impossible or highly unlikely values, such as wind directions of more than 360 degrees or soil temperatures greater than 45°C. However, as seen on figure 8a), some bogus data could not be removed with this method, as the spurious values still fell within a physically plausible range. Instead, filtered datasets were manually inspected to remove such unrealistic discontinuities.

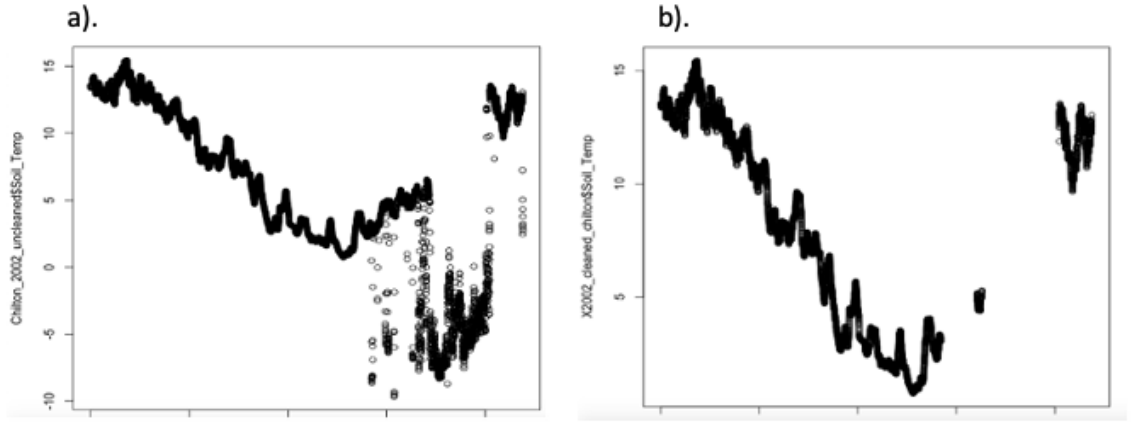

Figure 8: *Raw and cleaned soil temperature data for Chilton station at different periods through the cleaning process. Figures generated in R Studio.*

2.3 Imputation Model: Rain

Rainfall data was collected from several weather stations around the study area. Some datasets were excluded if they were extremely incomplete and closely located to a more complete station or contained large quantities of spurious data. Unfortunately, the Chilton rainfall dataset contained many potential errors and discontinuities in its dataset, so was excluded. Three rain gauges in Arthur's Pass village are available, however only the ECan rain gauge provides a full record from 1955 to present, thus only it was selected for use within the model. The Rainfall records for the two rain gauges located at Cass Station were combined. A map (figure 9a) and the continuity of the datasets used in the rainfall model (figure 9b) are shown below.

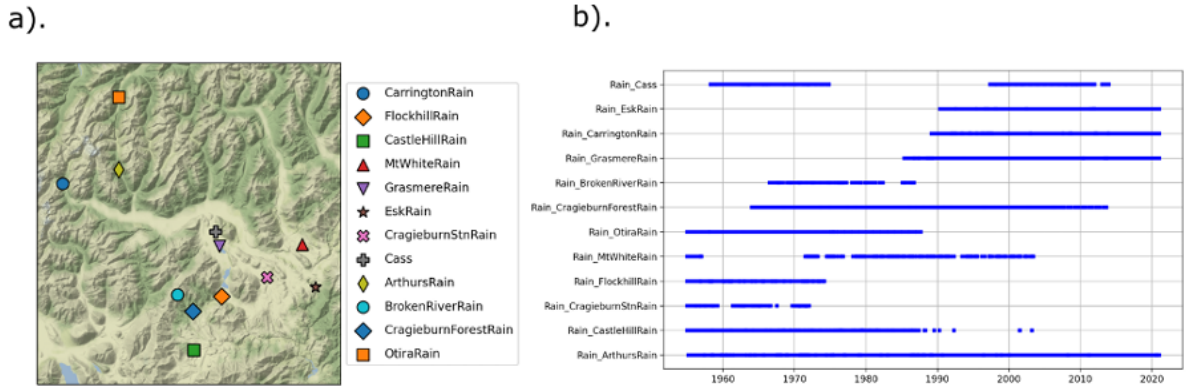

Figure 9: *a). A map of rain gauges in Arthurs Pass and b). a plot of their availability.*

A major problem presented by the datasets is that some rainfall gauges record daily rain at 0000hrs and others at 0900 or 0800 hours. This causes anywhere from 16 to 18 out of 48 total hours of recorded rainfall difference between two stations recording at these different times. These differences between daily rainfall totals cannot be easily corrected. Instead of daily rainfall, weekly rainfall totals were used, as this reduced amount of time difference between any two stations to only 18 of 336 total hours.

One issue with RF imputation of rainfall data is that the voting scheme makes it sometimes difficult for the imputer to reproduce weeks where there are no rainfall. To remedy this, a binary dummy variable was imputed alongside the rainfall date with 0 corresponding to a no-rain observation. This is then multiplied with the corresponding rainfall column thus setting any weeks that the model predicted had no rain are set to 0.

2.4 Imputation Model: Temperature

Temperature data is scarcer than rainfall due to the low number of weather stations in Arthur's Pass that record daily temperatures. Most weather stations, especially older stations, do not record daily mean temperatures, so only maximum and minimum temperatures were imputed. The locations and dataset spans are shown in figure 10 below. In the end, six stations were used: Arthurs Pass, Arthurs Pass EWS, Broken River, Craigieburn Forest, Chilton, and Cass. Because of its proximity and the limited span Arthurs Pass EWS, it was used to fill a few small gaps in Arthurs Pass in an initial 'pre-imputation' phase and then dropped from the dataset to reduce overall missingness and redundancy.

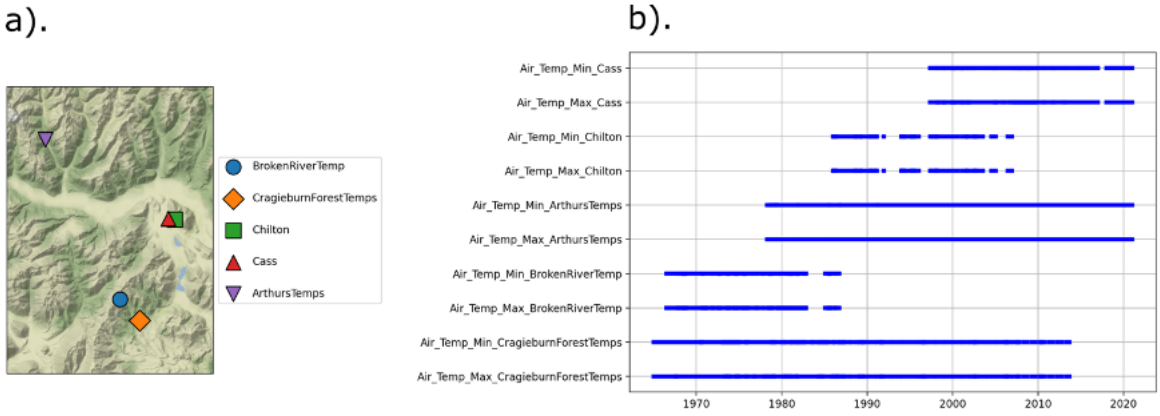

Figure 10: *a). A map of the weather stations recording temperature in Arthurs Pass and b). a plot of their availability.*

3. Results

3.1 Evaluating Imputations

The Datasets were split into testing and training portions for evaluation, this was achieved using two methods. The first method samples a fixed number of random "testing" from each variable across the span of the dataset. To avoid introducing spurious model uncertainty, the imputation algorithm was run multiple times using different sets of testing samples. In each case, 1000 total samples were taken with no individual dataset losing more than 0.5% of total data to the testing sample. This necessitated no less than 30 individual imputations for the rain data. This is advantageous because it accounts for the imputation accuracy of all variables across the entire dataset, however it less suitable for qualitatively evaluating the actual behaviour of an imputation. The second simulates a station blackout by removing all records from only one weather station over a time period. For hyperparameter optimization, random sample validation of all variables was used whereas simulated blackout validation was used to visually check the model and examine how imputation accuracy changed over time.

Models were evaluated according to R^2 scores, Mean Square Error (MSE), Mean Absolute Error (MAE), and Mean Bias (MB). R^2 records the proportion of variance in the test data that can be explained by variance in the predicted data. MSE and MAE are two measurements of the mean error across the dataset, MAE is the average magnitude of error whereas MSE is a measurement more sensitive to extreme outliers. MB is another average error measurement, but it considers the sign of the error, MB measures whether the model is consistently under or over predicting values and indicates whether lower resolution samples of the data, like annual or seasonal means, are likely to systematically overestimate or underestimate the true value of the data.

3.2 Rainfall

The hyperparameters were tuned by trial and error, it was found that a model utilising the following parameters yielded good results:

Table 1: Optimised hyperparameters for the forest regressor of the rain model

Rain Data Imputer:	**Dummy Variable Imputer:**
Number of Estimators: 400	Number of Estimators: 200
Max tree depth: 20	Max tree depth: 10
Number of Features: 7	Number of Features: 3
Max iterations: 20	Max iterations: 20
Minimum number of samples split node: 2	Minimum number of samples split node: 2
Minimum number of samples to split a leaf: 1	Minimum number of samples to split a leaf: 1

Using these parameters, the model produced satisfactory cross-validation scores as shown in figure 11 below.

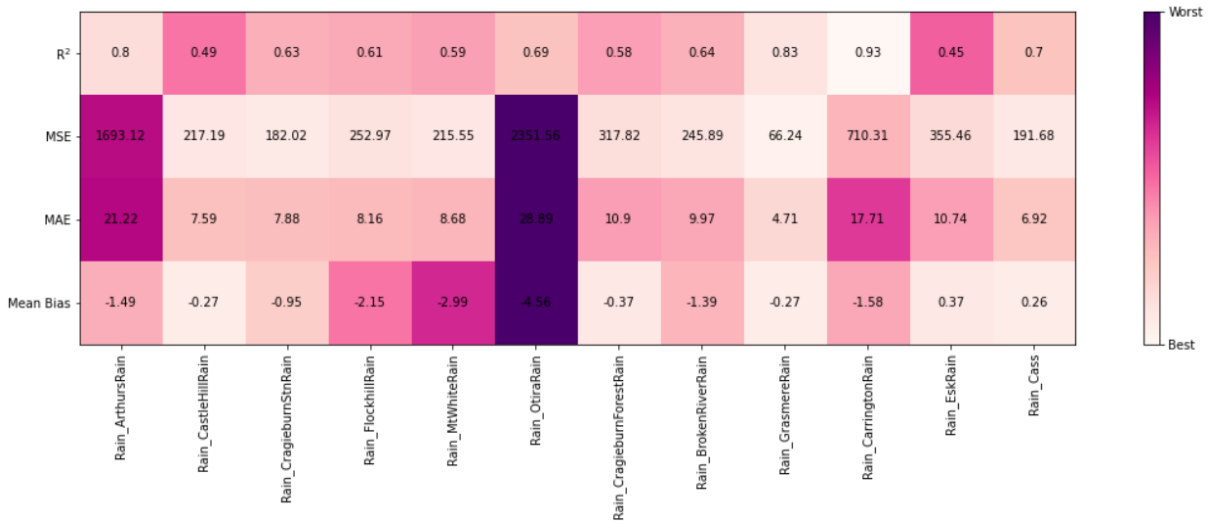

Figure 11: *Evaluation of imputation performance for rainfall data from ~1000 random samples taken from 30 imputations.*

Except for a few stations, error scores are generally low, however, high MSE error indicates that the model may be producing a number of outliers for some variables. The worst-performing stations are generally those found toward the edges of the dataset, which is likely to be due to the lower number of surrounding datasets those cadraw from. Comparing several test datasets from Cass it was found that imputation accuracy decreases in the older records. This is likely a result of the lack of closely located stations operating in the early dataset, unlike later on where the Grasmere rain gauge provides a very closely correlated dataset to the Cass station. However, it can be seen from Figure 12 a) below that the imputation still maintains a reasonably plausible record even in the older dataset with a nearly perfect record being produced post 1986 when Grasmere is in operation.

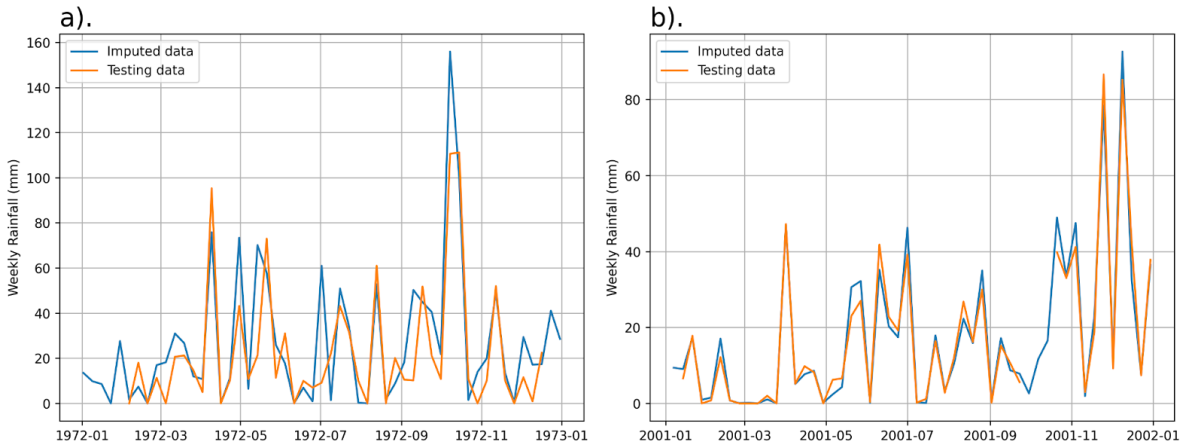

Figure 12: *a). imputed data filling a simulated station blackout for older data (1972) and b). imputed data filling a station blackout for 2001, after Grasmere station was in operation.*

3.4 Temperature

Hyperparameters for the temperature model were tuned using trial and error. The following hyperparameters (Table 2) were found to produce good results:

Table 2: Optimised hyperparameters for the forest regressor of the temperature model.

Temperature Data Imputer:
Number of Estimators: 500
Max tree depth: 20
Number of Features: 3
Max iterations: 20
Minimum number of samples split node: 2
Minimum number of samples to split a leaf: 1

Virtually all stations are imputed with relatively high correlation scores and low mean errors which indicate that the imputation is producing a satisfactory model in most cases (figure 13). The worst performance is found at the Chilton station and for minimum temperatures at Cass. The high MSE scores indicates that some outliers may be contributing to the inaccuracy of the imputation here, perhaps as the result of difficult-to-model local microclimatic effects. In any case, errors are still small and mean biases close to zero indicate that results are not systematically skewed, and therefore should produce reliable results at lower temporal resolutions.

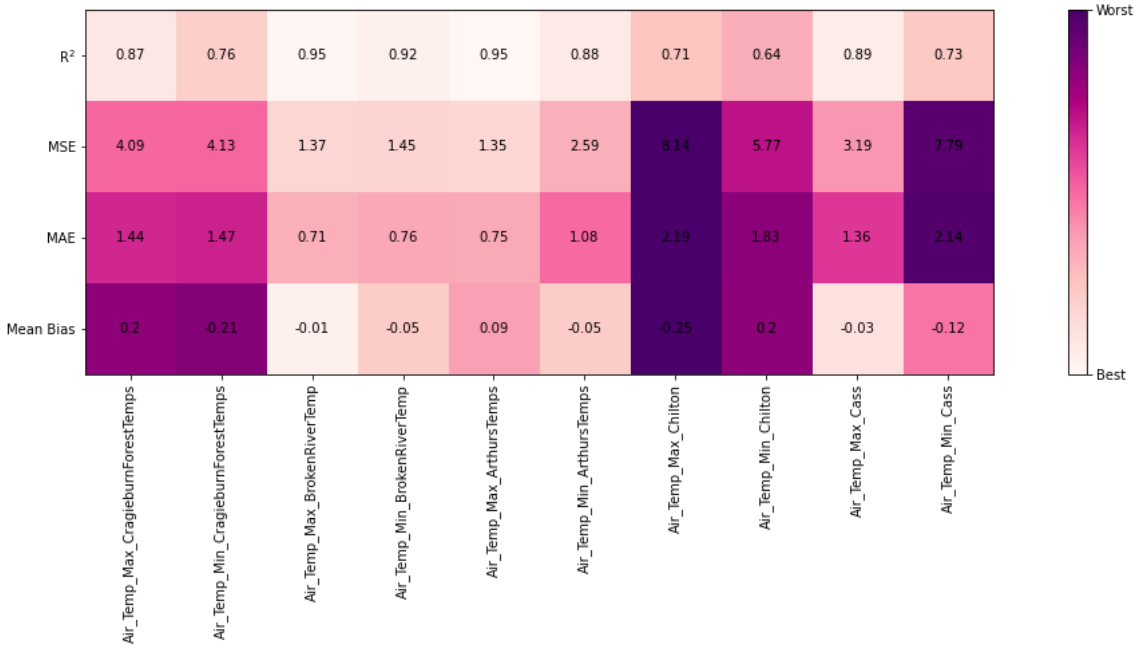

Figure 13: Evaluation of imputation performance for temperature data from ~1000 random samples taken from 30 imputations.

Two simulated station blackouts were introduced into the dataset, one in 1999 when the nearby Chilton temperature data is available (figure 14a) and in 2010 with no nearby stations (figure 14b). The timeseries shows an extremely plausible fit generated by the model in both cases.

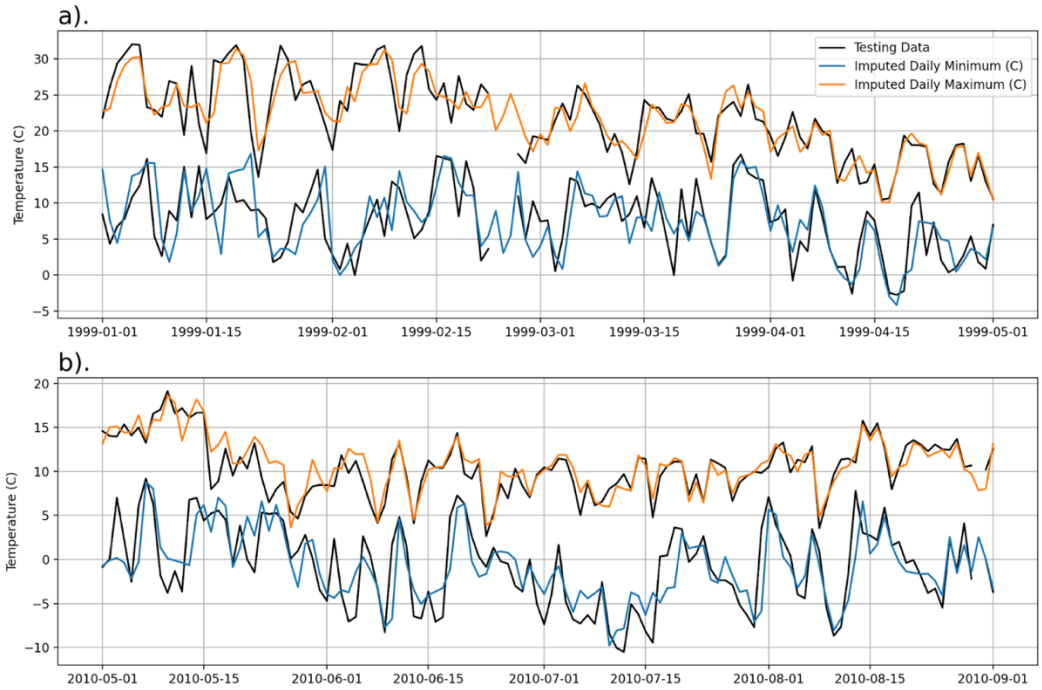

Figure 14: a). Simulated station blackout for 1999 with nearby Chilton data and b). Station blackout for 2010 with no available nearby stations.

3.5 Climate at Cass

Our dataset shows that the climate at Cass is typically warmer and drier than most of the other weather stations in the study area. Cass experiences substantially hotter mean maximum temperatures than any other station. Similarly, it also has the second highest mean daily maximum temperatures, second only to Chilton valley and similar to Arthurs Pass. High maximum temperatures might be explained by its relatively low altitude and exposure to sunlight year-round. Interestingly, Chilton does not appear to get as cold as Cass however, which might be due to the valley fog effect (1.6.2) and might explain Cass' overall similarity to Arthurs Pass, which is also located at the bottom of a valley.

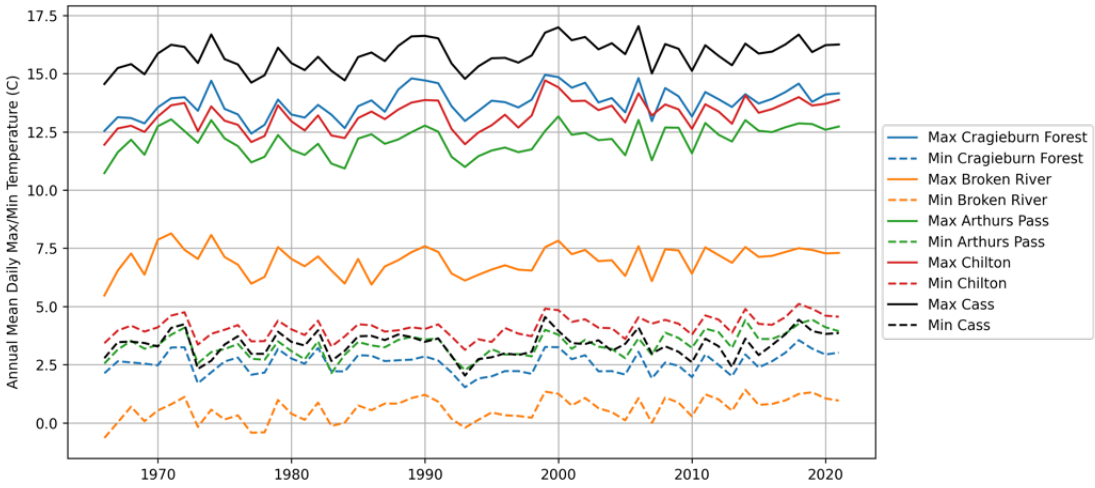

Figure 15: Imputed annual mean maximum and minimum temperatures.

Similarly to the temperature plot of figure 15 there is spatial variation in the volume of precipitation recordings (figure 16). It is clear that the higher precipitation recordings are from the weather stations that are located furthest west, thus experiencing more intense precipitation, compared to the eastern stations sheltered by the orographic rain shadow.

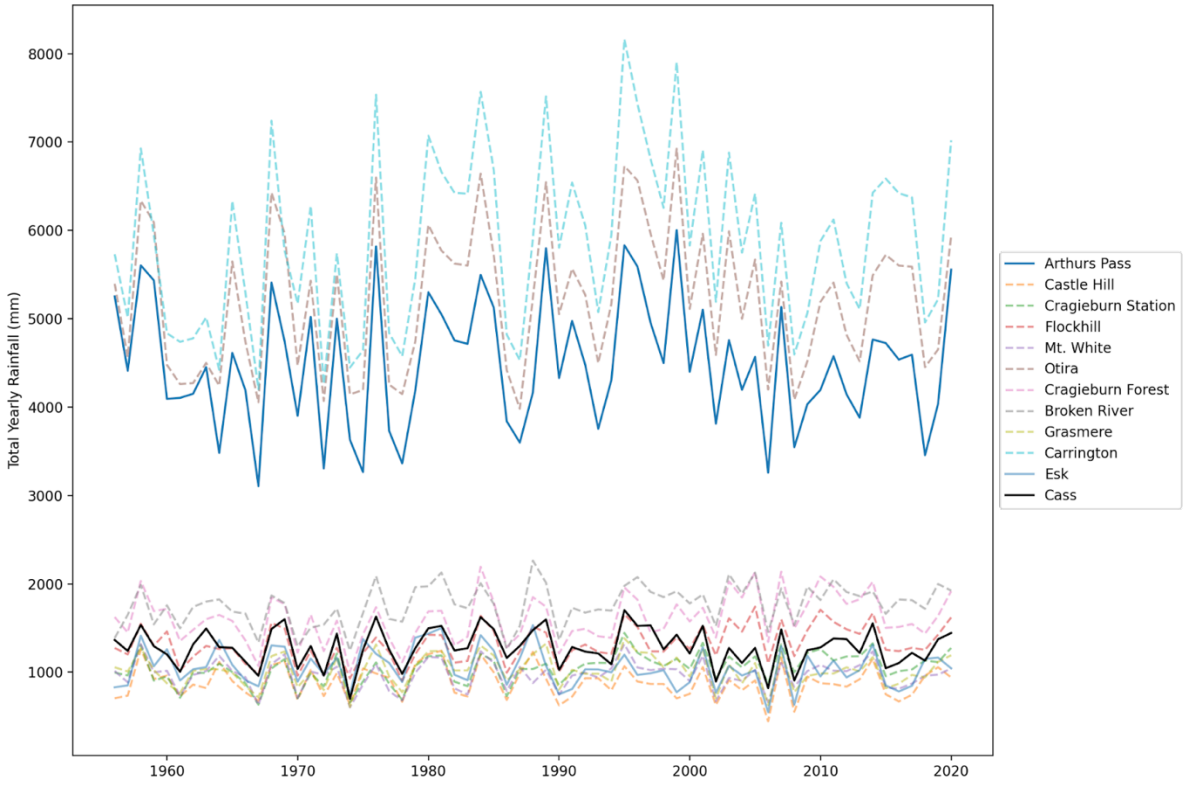

Figure 16: *Total annual imputed rainfall for all stations. Note the dramatic difference between west-coast and east-coast stations.*

4. Discussion

4.1 model success

Overall, both imputations for rainfall and Temperature appear to be satisfactory. These should not be treated as a perfect record; however, they are likely to be a good representation of the annual variability and an approximate account of weather at the stations. The most accurate Cass rainfall measurements will be those from 1986 onwards, with rainfall in this period being almost perfectly accounted for by the nearby rain gauge at Grasmere. The accuracy of Cass temperature estimates will be broadly similar across the dataset, although will probably be best represented from 1978 as due to the availability of Arthurs Pass temperature records, another station similarly located at the bottom of a valley and likely to experience similar micro-climates. Although this was a imputation mainly aimed at providing estimates for the climate at Cass, all datasets have been imputated. The accuracy of most these imputations are comparable to Cass, however caution is advised when using rainfall estimates from stations near the edge of the study area.

Together, the imputed datasets provide a fair accurate account of the general climate not only at Cass, but across the study area. Inter-annual trends are preserved across the dataset, and some climatic changes can be observed. For example, OLS regressing was used to fit a linear model for annual mean minimum and maximum temperature to Cass Station temperatures from 1965 to 2021 (table 4). We see a statistically significant positive trend in mean annual maximum temperatures of 0.0122^{o}C per year. This is in the order of the expected rate of change due to global warming, and it is interesting to note that this trend appears to be much weaker for minimum temperatures.

***Table 3:** Linear regression results for annual mean maximum/minimum temperatures at Cass.*

Variable	Coefficient (degrees/year)	P-value
Min Temp	0.0023	0.67
Max Temp	0.0122	0.032

4.2 Limitations

Although Random Forests makes no assumptions about the distribution or relationships between weather stations, it does make several important assumptions about the dataset itself. Because random forests are binary decision tress fitted to known observations, a key assumption they make is that the data provided is an accurate representation of the entire

dataset. In other words, it cannot extrapolate data into ranges that it has not seen before. This is potentially problematic if the range of a variable changes significantly over time. For example, since Cass data is only available from 1997, our imputation backwards until 1965 may be inaccurate due to the warming temperatures potentially allowing colder minimum temperatures in this period than were observed from 1997 onwards. Although limited evidence was found to suggest that this has affected our dataset, as Cass temperature trends closely follow trends observed at other stations with data in this time period, it remains an important limitation to consider when using very old, imputed data. Similarly, while the MissForest algorithm is resilient to large gaps found in our dataset (Zhu et al., 2022), large missingness will always introduce uncertainty in an imputation.

Finally, while we have done our best to ensure that the data used in the model was free of spurious data, it is possible that some bad readings slipped though. This is especially true for rainfall data that produces hard-to-catch errors (readings of 0mm) rather than clearly bogus data. Although the wisdom-of-the-crowd voting scheme and random bootstrap sampling of MissForest does make the algorithm much more resilient to outliers, they can still affect the stability and accuracy of the imputation.

5. Conclusion

In this report, we have used statistical and machine learning methods to create accurate estimates of missing weather station data at Cass station and surrounding areas in Arthurs Pass, New Zealand. We found that a Random Forest based imputation method (MissForest) (Stekhoven & Bühlmann, 2012) provided reliable and accurate results for both rain fall and temperature measurements. Using this method, we believe that we have significantly extended the range of reliable climate data for Cass basin and surrounding stations. This dataset, this report, and the open-source python code used in its production will be provided in a GitHub repository for whoever wishes to use it (see appendix A) .

Our hope is that this work will enable others to further study the climatological processes at Cass basin, to understand how they may be changing over time, and to inform decisions made by, landowners, Mana Whenua, policy makers and other stakeholders in the area. In future, the imputed datasets could be further extended to include more climate variables, with soil temperatures, solar readings, windspeeds, and humidity being obvious candidates. The results of this project has already been met with enthusiasm from potential stakeholders, including Mana Whenua, which is a clear indication of the need for such datasets. We believe the methodology pioneered in this project might also be used in other areas with similarly scant meteorological records.

Overall, this report highlights on the unique environment of Cass and Arthurs Pass, and all the challenges posed therein. We can see dramatic spatial variation across the pass and complex

weather patterns as a result of the mountainous landscape. We believe that our work will help connect people the processes that make this area so special and will guide the protection of this fragile ecosystem into the future.

6. Acknowledgments

We would like to thank Professor Peyman Zawar-Reza for supporting and directing us through this research project. We would also like to thank the Cass Research Area Management group represented by Dave Kelly for initiating this project, providing required data and answering any questions we had. We appreciate the support from Eva Nielson and David Evison whom expressed interest in our study. Additionally, Dr Abby Suszko for guiding through the process of developing Māori connections. Finally, we would like to thank the GEOG309 teaching team for making this project possible.

References

Aguilera, H., Guardiola-Albert, C., & Serrano-Hidalgo, C. (2020). Estimating extremely large amounts of missing precipitation data. Journal of Hydroinformatics, 22(3), 578–592. doi:10.2166/hydro.2020.127

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work?. International journal of methods in psychiatric research, 20(1), 40–49. https://doi.org/10.1002/mpr.329

Blandford, T. R., Humes, K. S., Harshburger, B. J., Moore, B. C., Walden, V. P., & Ye, H. (2008). Seasonal and synoptic variations in near-surface air temperature lapse rates in a mountainous basin. Journal of Applied Meteorology and Climatology, 47(1), 249-261.

Blandford, T. R., Humes, K. S., Harshburger, B. J., Moore, B. C., Walden, V. P., & Ye, H. (2008). Seasonal and synoptic variations in near-surface air temperature lapse rates in a mountainous basin. Journal of Applied Meteorology and Climatology, 47(1), 249-261.

Burrows, C. J. (1977). Cass: history and science in the Cass district, Canterbury, New Zealand.

Khanzode, K. C. A., & Sarode, R. D. (2020). Advantages and Disadvantages of Artificial Intelligence and Machine Learning: A Literature Review. International Journal of Library & Information Science, 9(1), 3. http://www.iaeme.com/IJLIS/issues.asp?JType=IJLIS&VType=9&IType=1

King, D. N. T., Skipper, A., & Tawhai, W. B. (2008). Māori environmental knowledge of local weather and climate change in aotearoa - new zealand. Climatic Change, 90(4), 385-409. https://doi.org/10.1007/s10584-007-9372-y

Lundquist, J. D., Pepin, N., & Rochford, C. (2008). Automated algorithm for mapping regions of cold-air pooling in complex terrain. Journal of Geophysical Research: Atmospheres, 113(D22).

Mucherino, A., Papajorgji, P.J., Pardalos, P.M. (2009). k-Nearest Neighbor Classification. In: Data Mining in Agriculture. Springer Optimization and Its Applications, vol 34. Springer, New York, NY. https://doi.org/10.1007/978-0-387-88615-2_4

Nowakowski, A. J., Frishkoff, L. O., Agha, M., Todd, B. D., & Scheffers, B. R. (2018). Changing thermal landscapes: merging climate science and landscape ecology through thermal biology.

Current Landscape Ecology Reports, 3(4), 57-72. https://link.springer.com/article/10.1007/s40823-018-0034-8

Overpeck, J. T., Meehl, G. A., Bony, S., & Easterling, D. R. (2011). Climate data challenges in the 21st century. science, 331(6018),700-702.

Perry, G. L., Sparrow, A. D., & Owens, I. F. (1999). A GIS-supported model for the simulation of the spatial structure of wildland fire, Cass Basin, New Zealand. Journal of Applied Ecology, 36(4), 502-518.

Pohl, B., Sturman, A., Renwick, J., Quénol, H., Fauchereau, N., Lorrey, A., & Pergaud, J. (2022). Precipitation and temperature anomalies over Aotearoa New Zealand analysed by weather types and descriptors of atmospheric centres of action. International Journal of Climatology, https://doi.org/10.1002/joc.7762

Potter, K. A., Arthur Woods, H., & Pincebourde, S. (2013). Microclimatic challenges in global change biology. Global change biology, 19(10), 2932-2939. https://doi.org/10.1111/gcb.12257

Qing-Ling, S., Xian-Feng, F., Yong, G., & Bao-Lin, L. (2015). Topographical effects of climate data and their impacts on the estimation of net primary productivity in complex terrain: A case study in Wuling mountainous area, China. Ecological informatics, 27, 44-54.

Rauika Māngai (2020) A Guide to Vision Mātauranga: Lessons from Māori Voices in the New Zealand Science Sector. Wellington, New Zealand.

Scoggins, A., & Fisher, G. (2002). Air pollution exposure index for New Zealand. New Zealand Geographer, 58(2), 56-64. https://doi.org/10.1111/j.1745-7939.2002.tb01635.x

Stekhoven, D. J., & Bühlmann, P. (2012). MissForest--non-parametric missing value imputation for mixed-type data. Bioinformatics (Oxford, England), 28(1), 112–118. https://doi.org/10.1093/bioinformatics/btr597

Sturman, A. P., Fitzsimons, S. J., & Holland, L. M. (1985). Local winds in the southern alps, New Zealand. Journal of climatology, 5(2), 145-1

Te Arawhiti The Office for Māori Crown relations. (N.d). Guidelines for engagement with Māori.

Te Ngāi Rūnanga o Ngāi Tahu. (2018). He rautaki mō te huringa āhua o te rangi Climate Change Strategy. https://ngaitahu.iwi.nz/wp-content/uploads/2018/11/Ngai-Tahu-Climate-Change-Strategy.pdf

Tranmer, M., & Elliot, M. (2008). Multiple linear regression. The Cathie Marsh Centre for Census and Survey Research (CCSR), 5(5), 1-5. http://hummedia.manchester.ac.uk/institutes/cmist/archive-publications/working-papers/2020/multiple-linear-regression.pdf

Wake, B. (2015). Global versus local. Nature Climate Change, 5(11), 974-974.
Wake, B. (2015). Global versus local. Nature Climate Change, 5(11), 974-974.

Whiteman, C. D., & Doran, J. C. (1993). The relationship between overlying synoptic-scale flows and winds within a valley. Journal of Applied Meteorology and Climatology, 32(11), 1669-1682.

Whiteman, C. D., Allwine, K. J., Fritschen, L. J., Orgill, M. M., & Simpson, J. R. (1989). Deep valley radiation and surface energy budget microclimates. Part I: Radiation. Journal of applied meteorology

Wolfensberger, D., Gabella, M., Boscacci, M., Germann, U., & Berne, A. (2021). RainForest: a random forest algorithm for quantitative precipitation estimation over Switzerland. Atmospheric Measurement Techniques, 14(4), 3169–3193. doi:10.5194/amt-14-3169-2021

Young, L. M., Norton, D. A., & Lambert, M. T. (2016). One hundred years of vegetation change at Cass, eastern South Island high country. New Zealand Journal of Ecology, 40(3), 289-301.

Zhang, G., Bateni, S. M., Jun, C., Khoshkam, H., Band, S. S., & Mosavi, A. (2022). Feasibility of Random Forest and Multivariate Adaptive Regression Splines for Predicting Long-Term Mean Monthly Dew Point Temperature. Frontiers in Environmental Science, 10. doi:10.3389/fenvs.2022.826165

Zhu, S., Clement, R., McCalmont, J., Davies, C. A., & Hill, T. (2022). Stable gap-filling for longer eddy covariance data gaps: A globally validated machine-learning approach for carbon dioxide, water, and energy fluxes. Agricultural and Forest Meteorology, 314, 108777. https://doi.org/10.1016/j.agrformet.2021.108777

Appendix A

https://github.com/CaJaDav/Cass-Weather-Station-Imputation?fbclid=IwAR3MQRGhYTrkso_1Rbs2hEU01jBIlUK6Ax5NKRu3zuWn86Cd24-JyhcNXqc

Appendix B

R Studio code for figures

```
#set the working directory
setwd("~/309 project")
#import and read in dataset
Cass_Hourly <- read_csv("~/309 project /Cass_Hourly.csv")
Chilton_Hourly <- read_csv("~/309 project /Chilton_Hourly.csv")
Cass_Daily <- read_csv("~/309 project /Cass_Daily.csv")
Chilton_Daily <- read_csv("~/309 project /Chilton_Daily.csv")


#import openair package
library(openair)


# index by timestamp instead of number
Chilton_Hourly$date <- as.POSIXct( Chilton_Hourly$date, format="%d-%m-%Y%H:%M" )


#change the colum name time to date
colnames(Chilton_2006_clean)[1]='date'

print( TestData$Time )
print( class(TestData$Time) )

#windrose
windRose(Chilton_Hourly, type = "season", wd="Wind_Dir", ws = "Wind_Speed",
       breaks=c(0,2,4,6,8),
       key= list(labels =c(">0-2",
                           ">2-4",
                           ">4-6"),
             hemisphere = "southern"))

windRose(CassHourly, type = "season", layout = c(4, 2))

#polarannlus plot
polarAnnulus(CassHourly, wd="Wind_Dir", ws="Wind_Speed", pollutant = "Air_Temp")

#polar plot
polarPlot(
 Cass_Hourly,
 pollutant = "Air_Temp",
 x = "Wind_Speed",
 wd = "Wind_Dir",
 hemisphere = "southern",
 title="Chilton 1993")

#plot difference between the incoming solar radiation at Cass and Chilton
plot(Cass_1999_Hourly$Li200)
plot(Chilton_1999_Cleaned_SJ$Li200)
```

```
#calendar plot
calendarPlot(Cass_2006_raw, pollutant = "Air_Temp")
calendarPlot(Cass_Hourly, pollutant = "Cass_Hourly$Air_Temp", annotate = "Cass_Hourly$Wind_Speed")
calendarPlot(CassHourly, year = "1999", pollutant = "Wind_Speed",
        breaks = c(0, 5, 10, 15,20))

#compare plots of cleaned and uncleaned data
plot(Chilton_1993_filtered$Air_Temp)
plot(Chilton_1993_clean$Air_Temp)
```

Appendix C

This is a complete database for all the resources I have collected from NIWA, ECan and the University weather stations.

---------- DESCRIPTION OF DIRECTORIES ----------

- Raw_Data:

This contains the 'raw' files in their (mostly) original formatting. These have been minimally altered by myself
and are not all that usable, i.e. different headers, formatting, timestamps, etc.

- Formatted_Data:

This contains the raw data in a consistent format. The Timesteps are regular, variables are consistent. Each
station has its own folder. Data is stored in .csv files by year.

- Filtered_Data:

This is a directory containing data that filtered to remove 'obviously bad' data, i.e. values outside of a
physically feasible range.

- Cleaned_Data:

This is a special directory contained data that has been curated by hand to remove obviously spurious, but hard
to contain data.

- SUGGESTED USE:

Generally, the formatted datasets can be used for most of the data provided by NIWA or ECan. Filtered datasets, or cleaned
datasets are strongly recommended for Chilton and Cass due to the high amount of bad data.

---------- COLUMN DICTIONARIES AND FORMATTING NOTEBOOKS----------

Much of the formatting of the datasets have been accomplished withing the Jupyter notebooks attached. In broad strokes,
these contain the python scripts used to create the Cleaned, Filtered, and Formatted datasets from Raw_Data. Because these

scripts are bespoke for each differently formatted datasource, I felt it would largely be a wasted effort to turn these in
to a complete library or executable.

The Daily_Column_Dictionary and Hourly_Column_Dictionary files contain a lookup table of all the known variable names
across the raw datasets used. The top row are the names used in the rest of the dataset. A full breakdown of variables can
be found below.

Changing the vairable names is as simple as editing the .csv and running the Notebooks again.

****
---------- DAILY VARIABLE NAMES ----------
****

Output_Code
Year
Day ---------- These are largely unused, just a product of the datalogger at Cass and Chilton

Time ---------- Time in 'YYYY-mm-dd' formatting

Li190
Li190_Max
Li190_Max_Time
Li200
Li200_Max
Li200_Max_Time ---------- Li190 and Li200 are solar sensors for photolysis-active and total radiance respectively.

Air_Temp
Air_Temp_Min
Air_Temp_Min_Time
Air_Temp_Max
Air_Temp_Max_Time ---------- Mean, min and max air temperature readings

5m_Air_Temp
12m_Air_Temp ---------- Air temperature at different heights (Chilton)

Soil_Temp

Soil_Temp_Min
Soil_Temp_Min_Time
Soil_Temp_Max
Soil_Temp_Max_Time ---------- Soil Temperatures (unspecified depth)

10cm_Soil_Temp
20cm_Soil_Temp
50cm_Soil_Temp
10cm_Soil_Temp_Min
10cm_Soil_Temp_Min_Time
20cm_Soil_Temp_Min
20cm_Soil_Temp_Min_Time
50cm_Soil_Temp_Min
50cm_Soil_Temp_Min_Time
10cm_Soil_Temp_Max
10cm_Soil_Temp_Max_Time
20cm_Soil_Temp_Max
20cm_Soil_Temp_Max_Time
50cm_Soil_Temp_Max
50cm_Soil_Temp_Max_Time -------- Soil Temperatures at specified depth

Ground_Temp
Ground_Temp_Min
Ground_Temp_Min_Time
Ground_Temp_Max
Ground_Temp_Max_Time ---------- Surface Temperatures

Rel_Humidity
Rel_Humidity_Min
Rel_Humidity_Min_Time
Rel_Humidity_Max
Rel_Humidity_Max_Time ---------- Relative Humidity (%)

10cm_Soil_Moisture
20cm_Soil_Moisture
50cm_Soil_Moisture
10cm_Soil_Moisture_Min
10cm_Soil_Moisture_Min_Time
20cm_Soil_Moisture_Min
20cm_Soil_Moisture_Min_Time
50cm_Soil_Moisture_Min
50cm_Soil_Moisture_Min_Time
10cm_Soil_Moisture_Max
10cm_Soil_Moisture_Max_Time
20cm_Soil_Moisture_Max
20cm_Soil_Moisture_Max_Time
50cm_Soil_Moisture_Max

50cm_Soil_Moisture_Max_Time --- Soil Moisture Readings

Wind_Speed