POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA CIVILE,
AMBIENTALE E TERRITORIALE

EXECUTIVE SUMMARY OF THE THESIS

Using Machine Learning to reconstruct temperature and precipitation climatologies in an Italian citizen science weather station network

TESI DI LAUREA MAGISTRALE IN ENVIRONMENTAL AND LAND PLANNING ENGINEERING - INGEGNERIA PER L'AMBIENTE E IL TERRITORIO

AUTHOR: FRANCESCA RAMPINELLI

ADVISOR: ALESSANDRO CEPPI

CO-ADVISOR: GUIDO CIONI

ACADEMIC YEAR: 2023-2024

1. Introduction

It is widely acknowledged that the climate is undergoing a rapid and alarming transformation. Anthropogenic global warming has already warmed the planet: from the preindustrial period (1850-1900) to the present (2011-2020) the global average temperature over land has increased by 1.1°C [1]. If global warming continues at the current rate, it is projected that the global average temperature will reach a 1.5°C increase between 2030 and 2052 [1].

Awareness of the ongoing changes allows for more effective preparation to address the challenges posed by climate change and to implement appropriate strategies for mitigation and adaptation. For this reason, datasets of monthly climatological normals of meteorological variables (or climatologies) at a high spatial resolution have proved to be of increasing importance in the recent past, and they are likely to become even more important in the near future [2]. Indeed, they are crucial in a variety of models and decision-supporting tools in a wide spectrum of fields such as agriculture, engineering, hydrology, ecology and natural resource conservation [2].

In recent years, Italy, like other countries in the Mediterranean area, has been observing an increase in extreme weather events (i.e., heavy rainfall and increase in temperature), causing huge impacts (i.e., floods, droughts, heat waves) with consequences to assets and people [3]. Moreover, their occurrence in multiple climatic impact-drivers perspectives is expected to increase (with high confidence) since global warming is expected to increase faster over this area than the global mean temperature change [1]. Therefore, there is an increasing need for more reliable and detailed climate information on the Italian peninsula to improve the assessment of climate hazards and risk management [4].

Given that a decline in operating ground-based weather stations has been observed in recent years in most countries around the world and that the demand for real-time, high space-time resolution data is increasing, the importance of crowdsourcing weather data becomes clear.

The Meteonetwork (MNW) system is a prime example of a citizen weather station network (CWS), covering a wide territory with high spatial density, which allows for a high redundancy of measurements.

2. Aim of the Study

This study aims to develop a methodology for reconstructing climatologies for the period 1991-2020 using daily data of maximum, minimum, and average temperatures, as well as cumulative precipitation, from the meteorological stations of the Citizen Science (CS) network Meteonetwork (MNW).

Given that the MNW association was established in 2002, there is a deficit of data from its station network prior to that year. In order to reconstruct climatologies, which require a minimum of 30 years of data, it is necessary to identify a method for obtaining the missing data.

The idea is to utilise the nearest grid points obtained from a gridded reanalysis model to establish a relationship between reanalysis and MNW station data over a specified period (reference), through machine learning (ML) techniques. This represents the first instance of ML models being applied to the MNW network with the specific objective of reconstructing climatologies. Once the relationship has been identified, it can be employed to reconstruct an equivalent synthetic time series for each station over the climatological period 1991-2020, which is the 30-year reference period established by the World Meteorological Organization (WMO).

3. Material and Methods

3.1. Study Area

The geographical area of study is Italy, situated in the centre of the Mediterranean region, with an area of approximately 302,000 km².The country's geographical position, coupled with the presence of mountain ranges such as the Alps and the Apennines, gives rise to considerable climatic differences between the various regions. The climate varies considerably across Italy, from the temperate continental climate of the northern areas, with cold winters and hot summers, to the Mediterranean climate of the southern coastal areas and island, characterised by hot, dry summers and mild, rainy winters. In contrast, mountainous regions experience lower temperatures and precipitation during the winter snowfall season.

3.2 Data

Of the 4,411 meteorological stations comprising the MNW network in Italy, 2,163 are managed by citizen scientists. In order to identify an effective methodology, 43 stations of the aforementioned group with the longest time series of the variables of interest were selected from the MNW database (Figure 1).

The historical time series for each variable of interest (T avg, T max, T min and P) of the selected stations range from 9 to 23 years, in Figure 2 are shown the observation periods of precipitation.

It is therefore necessary to identify a method for reconstructing the time series of the stations in order to ensure the availability of at least 30 years of observations for the calculation of climatologies.

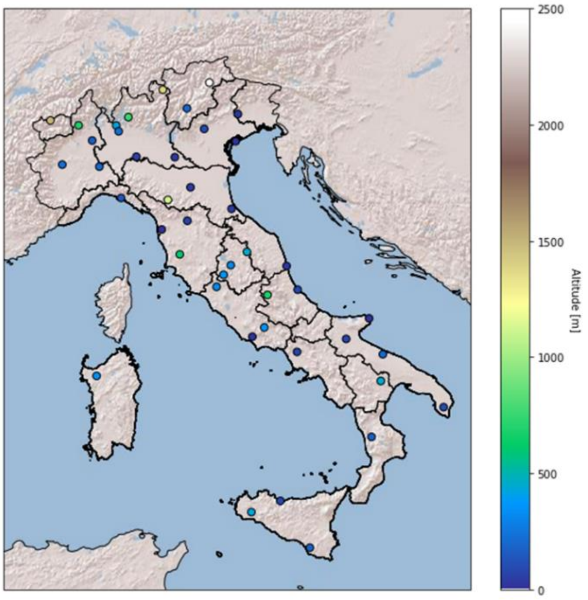

Figure 1: Selected MNW observation station, their location and altitude in meters.

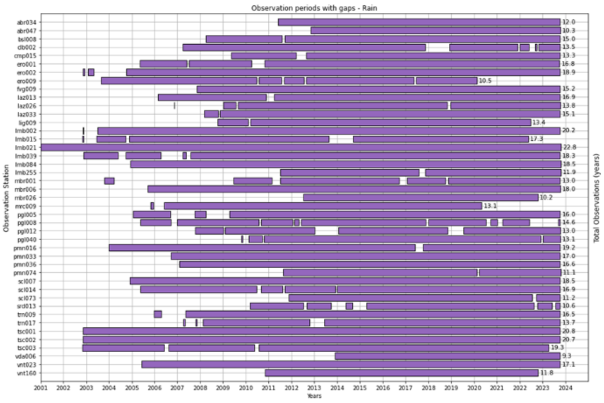

Figure 2: Observation periods of variable precipitation and effective length of historical series of each station in years.

Furthermore, a number of reanalysis datasets were selected for testing, in order to ascertain which one should be chosen for input into the ML procedure. Five different reanalysis datasets were analysed for temperature, two of which were global and three regional. The global datasets were Era5-Land and MSWX, while the regional ones were CERRA, MERIDA H_RES and VHR_REA_IT. In addition to the aforementioned reanalyses, two further precipitation datasets were considered: MSWEP and CHIRPS.

All data series of daily maximum, average and minimum temperature and daily cumulative precipitation of the grid points of each reanalysis surrounding each station were extracted already aggregated on a daily scale from 1991 to 2020.

3.3 Methods

Two types of machine learning (ML) models were also selected, which have previously demonstrated efficacy in similar contexts: Support Vector Regression (SVR) and Random Forest (RF).

In order to obtain a model that reconstructs the MNW station time series as accurately as possible, it is essential to select a suitable input, and therefore it is essential to select a reanalysis that has a high degree of compatibility with the observations recorded at all stations. To this end, each reanalysis dataset is compared with all the observations from all 43 stations together, first by comparing the raw reanalysis data and then by comparing the reanalysis data corrected with simple procedures that take into account the distance of the grid points from the station, by interpolation with IDW, and the differences in altitude between the grid points and the stations. These corrections were made because global and regional reanalyses often show systematic and regionally distributed biases with respect to observations from meteorological ground stations. Subsequently, for each variable and station, all potential combinations between the distinct model types (SVR and RF) and the selected reanalysis datasets were evaluated to determine the optimal performance, with varying values of the models' hyper-parameters. This was achieved through a Grid Search procedure to optimise the parameter values and k-fold cross-validation to prevent overfitting of the models.

4. Results and Discussion

For temperature, the reanalysis data set with the best agreement with observations for all temperature variables was CERRA (Figure 3). It was therefore decided to continue the analysis using this dataset. In addition, Era5-Land was chosen as a complementary analysis as it has large temperature time series extending to mid-2023, a period two years longer than the time series available for CERRA. The aim was to test whether this could be a factor influencing the performance of the machine learning models. ERA5-Land is also the “runner-up” reanalysis after CERRA, when elevation correction is applied.

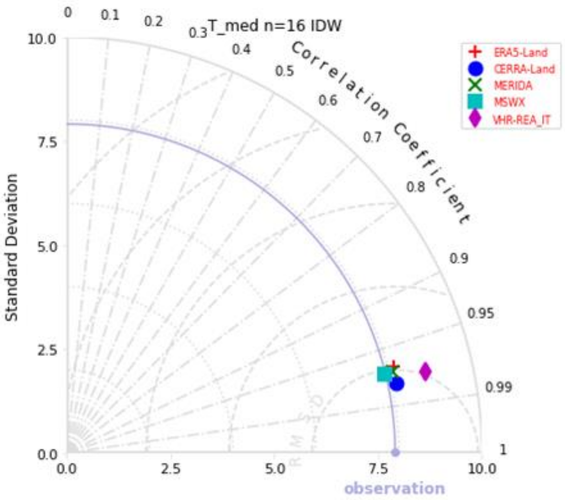

Figure 3: Taylor diagrams of the variable T avg with each reanalysis dataset compared with all observed data from all 43 stations collectively, considering only the nearest 16 grid points, aggregated with IDW.

In order to assess the models' performance in simulating precipitation, the following three models were selected: ERA5-Land, CERRA and MSWEP. The MSWEP model demonstrated the most optimal performance (Figure 4), exhibiting the highest correlation coefficient and the lowest RMSE. The CERRA and ERA5-Land models, which exhibited a relatively close standard deviation to the observed data, were selected as the second and third best models, respectively.

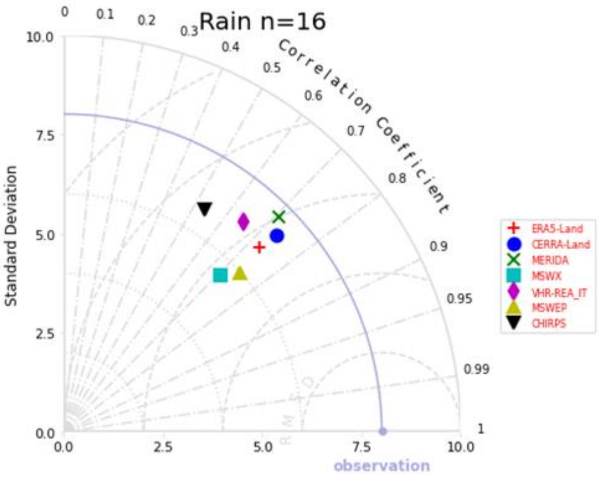

Figure 4: Taylor diagrams of the variable P with each reanalysis dataset compared with all observed data from all 43 stations collectively, considering only the nearest 16 grid points, aggregated with arithmetic mean.

With regard to temperature, all the tested models demonstrated high accuracy, with R^2 values exceeding 0.85 and RMSE values consistently below 2°C in validation for all the stations and the variables. The models exhibited an enhanced aptitude for representing **mean temperature**, achieving the highest **R^2** values (between **0.96** and **0.99**) and the lowest **RMSE**s, which were equal to or less than **1.5°C**. In contrast, the reconstruction of **minimum temperature** exhibited the poorest performance, with **R^2** values between **0.85** and **0.98** and **RMSE** values between **1** and **2.5°C**. On the other hand, **maximum temperature** demonstrated a more favourable outcome, with **R^2** values between **0.90** and **0.98** and **RMSE** values between **1** and **2°C**. The two models, Support Vector Regression (SVR) and Random Forest (RF), exhibited comparable and highly analogous performance when the optimal parameters for each station were identified. However, **SVR with linear kernel** demonstrated slight superiority in the majority of cases. The latter exhibited optimal performance in 67% of instances for maximum temperature, 81% for mean temperature and 84% for minimum temperature.

Similarly, the two reanalyses (CERRA and ERA5-Land) demonstrated comparable performance when employed as inputs to the same model, with minimal discrepancy.

The models employed for the reconstruction of **precipitation** series were considerably more complex than those utilised for temperature series. The incorporation of **sample weights** into the models proved to be a crucial factor in enhancing their predictive capacity. In the absence of such weights, the models were unable to identify an optimal function, resulting in the generation of negative R^2 values in some cases. The optimal weight value, which yields the best performance, is **10^4** for days with precipitation and zero for days without precipitation. The values of the **coefficient of determination** demonstrate considerable variability in the process of validation, with a range of **0.1** to **0.8** depending on the specific station under consideration. In the case of 11 stations (26% of the total), it was not possible to obtain a model that exceeded the performance of R^2 equal to 0.5 in the validation process. In contrast, the **RMSEs** demonstrate a relatively consistent error, varying between approximately **3 and 6 millimetres**. In contrast to temperature, which exhibited minimal variability, the efficacy of precipitation models is contingent upon both the specific model employed and the reanalysis data utilised as input.

The only model that demonstrated an acceptable level of accuracy for all observation stations was the **SVR with a radial (RBF) kernel.** In contrast, the use of SVR with a linear kernel required an excessive amount of computational time for the Grid Search method to identify optimal hyper-parameters, rendering the process impractical on a standard laptop within a limited timeframe. While Random Forests demonstrated favourable outcomes at certain stations, they exhibited inferior performance compared to SVR and were effective only at a very restricted number of stations. The most effective reanalysis for precipitation was **MSWEP**, which demonstrated optimal performance in 77% of cases.

For these reasons, the temperature series were reconstructed using support vector regression (SVR) with a linear kernel and the CERRA reanalysis, whereas the precipitation series were reconstructed using SVR with a radial kernel and the MSWEP reanalysis, based on the average performance obtained at all stations.

An example of the resulting climatologies is shown in Figure 5.

Finally, in order to verify the accuracy of the reconstructed climatologies, a comparison was conducted between these and the climatologies calculated using only the daily observations of the available stations, despite the presence of missing data, during the period of overlapping. Although this comparison is not an exact reflection of the actual data, as the calibration and validation of the models occurred during the same period, it provides a means of determining, in addition to the model accuracy indicators, whether the reconstructed series are sufficiently similar to the actual series.

With regard to temperature, it can be observed that even under the most unfavourable circumstances, the monthly climatologies show a minimum deviation of a few degrees from the actual data, thereby demonstrating their effectiveness as a representation of climatological averages (Figure 6). It can, therefore, be reasonably assumed that the results will be favourable even in periods where there is no overlap, due to the good accuracy of the obtained models.

With regard to precipitation, a comparison of the best station climatologies reveals that the cumulative monthly averages of precipitation exhibit a relatively low degree of divergence from the actual recorded values. The greatest divergence is observed in April, with a difference of approximately 7.5 mm (11.5%). The remaining months always exhibit a discrepancy of less than 6 mm, with an average deviation of 5.8%. It can be also observed that the model underestimates precipitation in half of the months, particularly during the winter months, while overestimating it in the summer months.

For the station exhibiting the poorest model performance (Figure 7), there is a discrepancy of about 20 mm (39%) between the climatologies for the initial six-month period. It is therefore evident that the model is unsuitable for accurately representing the climate trend of the station.

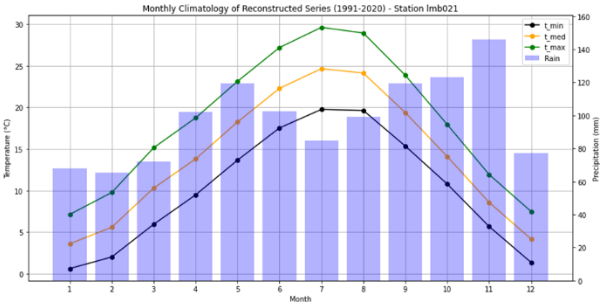

Figure 5: Monthly climatologies of the period 1991-2020 of the reconstructed synthetic series of station lmb021.

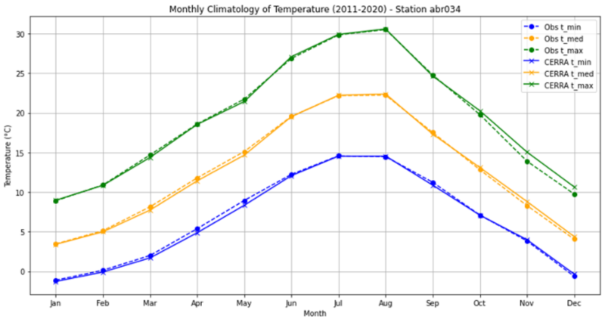

Figure 6: Comparison between monthly reconstructed climatologies and the observed ones during the common period, for station abr034.

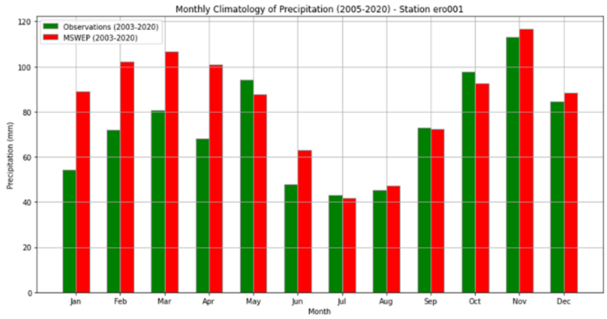

Figure 7: Comparison between monthly climatologies of precipitation reconstructed and the ones obtained from the observation on the common period for station ero001.

5. Conclusions

It was thus demonstrated that the designed methodology was highly effective in the reconstruction of temperature series, thereby yielding reliable climatologies in instances where complete historical data were unavailable. However, the performance for precipitation was more variable, with the accuracy of the model dependent on the station and the input data. It seems probable that this discrepancy can be attributed to issues pertaining to the measurement of observed station data. An alternative explanation may be found in the absence of heated rain gauges at numerous stations. Furthermore, even the regional reanalysis datasets may be unable to resolve localised phenomena such as thunderstorms, which is likely to have contributed to the poorer performance of precipitation models. The MSWEP dataset had a resolution of about 10 km, which was insufficient for a number of stations. This highlights the need for more detailed input data in future analyses. Moreover, precipitation exhibits markedly disparate temporal and spatial gradients when compared to temperature. This renders the training of an ML model for precipitation more challenging. In the absence of the sample weights, the model is unable to discern the relationship between the reanalysis data and observations.

In conclusion, this approach represents a promising solution to the problem of incomplete observational data, enabling the reconstruction of climatological time series. Nevertheless, the efficacy of this methodology is contingent upon the quality of both the station observations and the reanalysis data employed. In particular, further advancements in reanalysis resolution and model capabilities will be essential to improving the accuracy of precipitation reconstruction in areas where localised weather phenomena play a significant role.

References

[1] K. Calvin et al., «IPCC, 2023: Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland.», Intergovernmental Panel on Climate Change (IPCC), lug. 2023. doi: 10.59327/IPCC/AR6-9789291691647.

[2] C. Daly, W. Gibson, G. Taylor, G. Johnson, e P. Pasteris, «A knowledge-based approach to the statistical mapping of climate», Clim. Res., vol. 22, pp. 99–113, 2002, doi: 10.3354/cr022099.

[3] M. C. Llasat et al., «High-impact floods and flash floods in Mediterranean countries: the FLASH preliminary database», Adv. Geosci., vol. 23, pp. 47–55, mar. 2010, doi: 10.5194/adgeo-23-47-2010.

[4] M. Adinolfi, M. Raffa, A. Reder, e P. Mercogliano, «Investigation on potential and limitations of ERA5 Reanalysis downscaled on Italy by a convection-permitting model», Clim. Dyn., vol. 61, fasc. 9–10, pp. 4319–4342, nov. 2023, doi: 10.1007/s00382-023-06803-w.

[5] M. Brunetti, M. Maugeri, T. Nanni, C. Simolo, e J. Spinoni, «High-resolution temperature climatology for Italy: interpolation method intercomparison», Int. J. Climatol., vol. 34, fasc. 4, pp. 1278–1296, mar. 2014, doi: 10.1002/joc.3764.

[6] A. Crespi, M. Brunetti, G. Lentini, e M. Maugeri, «1961–1990 high-resolution monthly precipitation climatologies for Italy», Int. J. Climatol., vol. 38, fasc. 2, pp. 878–895, feb. 2018, doi: 10.1002/joc.5217.

[7] M. Giazzi et al., «Meteonetwork: An Open Crowdsourced Weather Data System», Atmosphere, vol. 13, fasc. 6, p. 928, giu. 2022, doi: 10.3390/atmos13060928.