Imputing Gaps in Swiss River Dataset

Bachelor Thesis
Faculty of Science, University of Bern

submitted by

Carlo Robbiani

from Novazzano TI, Switzerland

Supervision:

PD Dr. Kaspar Riesen
Benjamin Fankhauser
Institute of Computer Science (INF)
University of Bern, Switzerland

Abstract

The Federal Office of the Environment (FOEN) of Switzerland has been collecting water temperatures and discharge data of the four Swiss rivers Rhine, Rhone, Inn and Ticino for over 40 years. Although the data collection was carried out carefully, gaps of different sizes occur in the data of the roughly 80 water stations. These gaps happen due to various reasons such as scheduled maintenance activities or sensor malfunctions, which pose significant challenges to comprehensive data analyses and the accuracy of predictive models. This paper aims to use several pretrained LSTM (Long Short-Term Memory) models, introduced in previous work, to impute the gaps in the water temperature, using advanced strategies to handle missing data in the input variables. In order to measure the performance of these strategies, the percentage of gaps imputed with each model is calculated. Not using any advanced strategies, proposed in this paper, sets a solid baseline, as already every gap in the dataset can be imputed, using the best model in 52% of the cases. When incorporating the more complex strategies, that aim to handle cases when input variables of the LSTMs are missing, the best performing model used can be applied to 89% of the gaps, achieving a state-of-the-art RMSE of 0.652 over all predictions.

Acknowledgements

I am grateful to my supervisor Benjamin Fankhauser for his constant support and constructive feedback in writing my Bachelor's thesis. I appreciate his openness for discussion at all times and his willingness to address any issues or concerns I had. Furthermore I also want to thank him for taking the time to hold regular meetings with me, to discuss my questions and the next steps to take. His feedback was helpful in giving my thesis its final polish. I am also grateful to Prof. Dr. Riesen for his expertise and the opportunity to write this thesis. His knowledge on academic writing was invaluable for writing the thesis.

Contents

Chapter 1

Introduction

This Chapter will give a brief introduction to the work, discuss the Goal of the paper the models that are used and the dataset. This Chapter will be concluded with a discussion about the different Gaps encountered in the data and the different challenges of these.

1.1 Topic

River temperature is an important factor that can be used to determine the health of an aquatic ecosystem. All aquatic species have a specific river water temperature range that they can tolerate and significant changes in river water temperature might detrimentally affect aquatic species [1]. Therefore, having data on the water temperature of rivers is important for gaining more insights and get a better understanding of the change in the water temperature during a longer period of time. This is particularly relevant when considering the effects of climate change that impact the river temperature [2], [3].

The lack of available long-term data on stream temperatures has been recognized as a major limitation for understanding thermal regimes of riverine ecosystems [4].

The Federal Office of the Environment (FOEN) [1] of Switzerland has been systematically gathering water temperature and discharge data of the Swiss rivers for over 40 years starting from 1980. However, the measurements of water temperature often contain gaps of varying sizes. These gaps can be caused by a multitude of factors, including scheduled maintenance, data logging errors, accidental damage to the equipment or clogged sensors. Such disruptions highlight the challenges in maintaining continuous and accurate water temperature records. The presence of missing data not only complicates the analysis but also affects the performance of

[1] https://www.bafu.admin.ch/bafu/en/home.html

predictive models, which rely on continuous and complete datasets for training and validation. Addressing these gaps through robust imputation methods can be seen as major data quality improvement and thus accelerate the research of more sophisticated models. In the scope of this project, we also have access to atmospheric data like the air temperature provided by MeteoSwiss's network of air temperature measuring stations.

1.2 Goals

I received access to the datasets of The Federal Office of the Environment of Switzerland that has been collecting water temperature data of the Swiss rivers for over 40 years. The data of the water temperature and the air temperature consists of the daily averages. Additionally to the water and air temperatures we use discharge measurements. The discharge measures how much water per time unit flows through a specific river segment, usually measured in m^3/s. Alongside these measurements, results from prior research done will also be used.

Prior literature introduces a graph structure to represent the connectivity between neighbouring water and air stations [5]. In further work several LSTMs (Long Short-Term Memory) are introduced, which leverage information from neighbouring stations, discharge and/or air temperature to be able to impute water temperatures with greater accuracy [6].

The goal of this work is to apply these LSTM models, introduced by Fankhauser et. al. [6], to impute the real world gaps in the water temperature dataset. As the more advanced LSTM models require more input variables, the presence of these gaps in the water temperature, air temperature or discharge measurements poses a significant challenge, as the models cannot be utilized when there is missing input. In order to address this issue and not only rely on the models with lower accuracy due to their simplicity, several strategies are introduced to handle missing input variables. By implementing these strategies this work aims to enhance the quality of the real world water temperature dataset and therefore providing the basis for more accurate and reliable hydrological analyses.

1.3 Dataset

This section will discuss the structure of the dataset, the different gap types and when they occur.

1.3.1 Different Gaps

During the analysis of the dataset I categorized the different gaps into these categories:

- Short gaps (two days or shorter)
- Long gaps (more than two days)
- Special gaps

The first gap category are the short gaps. These gaps include all the gaps which are two days or shorter, i. e. gaps of length one or two days. The long gaps include all the other gap lengths. These two types of gaps are differentiated, since for the short gaps interpolation can be used to get satisfying results. Whereas this would also be possible for the longer gaps, the imputation result would suffer a significant loss in accuracy. The gaps in those two categories do not show any missing data in their neighbours. The special gaps include all the cases, where it is not possible to use every LSTM models as is, since the more advanced models require as input air temperatures, discharge and water temperatures. The special gaps are divided into these three categories:

- Stations built between to existing stations
- Gaps with some missing neighbours
- Gaps with only missing neighbours

The first category of gaps occur when a station was built in between two already existing stations, therefore resulting in a gap in the neighbour.

The second category are gaps with some missing neighbours during the same time period. This means, that a station has at least one missing neighbour and also at least one neighbour that recorded data during this time period.

The third category are gaps that have only neighbours with missing values during that time period, meaning that such stations do not posses any neighbour with recorded values during this time period. This makes it particularly hard to impute them, as the better performing models require the data from these neighbours. The second and third case are distinguished, because two different strategies are used to deal with them. The second case, where there are no neighbours with recorded values, mostly comes from the fact, that some stations were only built several years later than their neighbours. The two most common cases are when a new river branch was equipped with water stations or a new station was built in between two

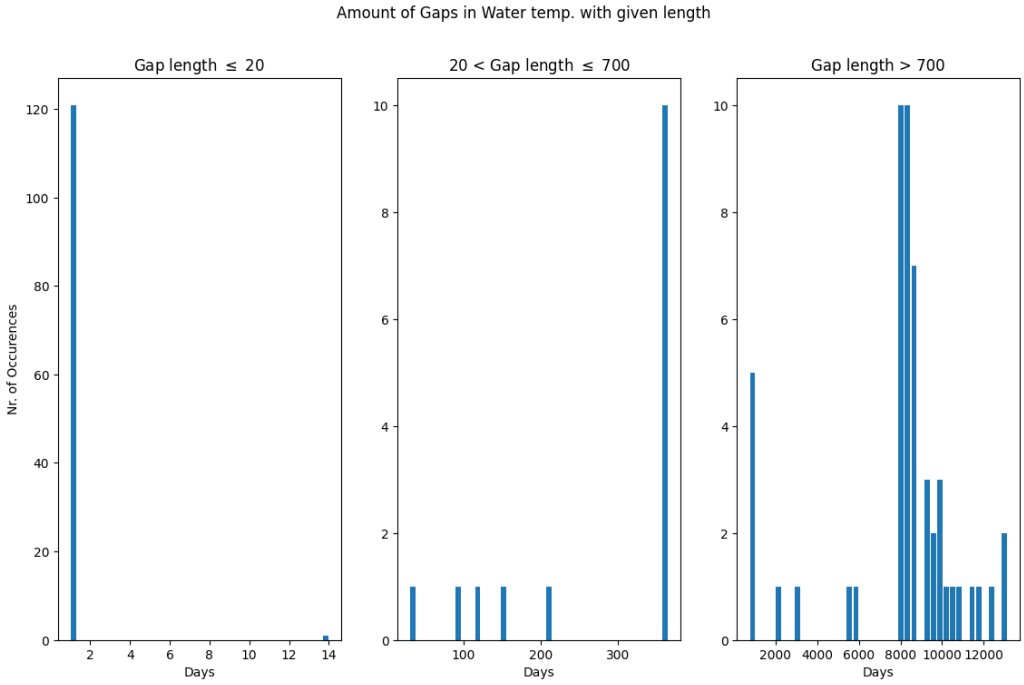


Figure 1.1: Nr. of occurrences of different gap lengths [6]

already existing stations. This leads to the situation where certain water stations do not have neighbouring data from every neighbour for more than 20 years. I will go into more detail on how these gaps are handled in Chapter 3.

In Fig. 1.1 you may see the distribution of the different gap occurrences [6]. As you may see most gaps are only one day long with about 120 of these. These short gaps may have occurred due to planned maintenance or sensor failure which prevented the sensor from recording temperature during this day. There are only a few gaps with the length of 20 up to 700 days as you may see in the middle graph. In the right graph are the gaps with a length of more than 700 days. The gaps with a gap length of more than 2000 days are mostly the aforementioned newly built stations. Most of these stations were only built between the years 2000 to 2010 and lack the data before their construction. Since I am also interested in the data from 1980 up until the station was built, we have a gap during this period which needs to be imputed, hence the long gaps. While these long gaps are not represented as often as the short gaps, they make up the majority of the missing values due to their length.

In the heatmap in Fig. 1.2 you see the distribution of the gaps for each water station in the years 1980 to 2020. A cell is colored dark grey if the station recorded data on that day and light grey if the station has missing values. As you can see, the long gaps from the roughly 40 newly built stations are mainly responsible for the

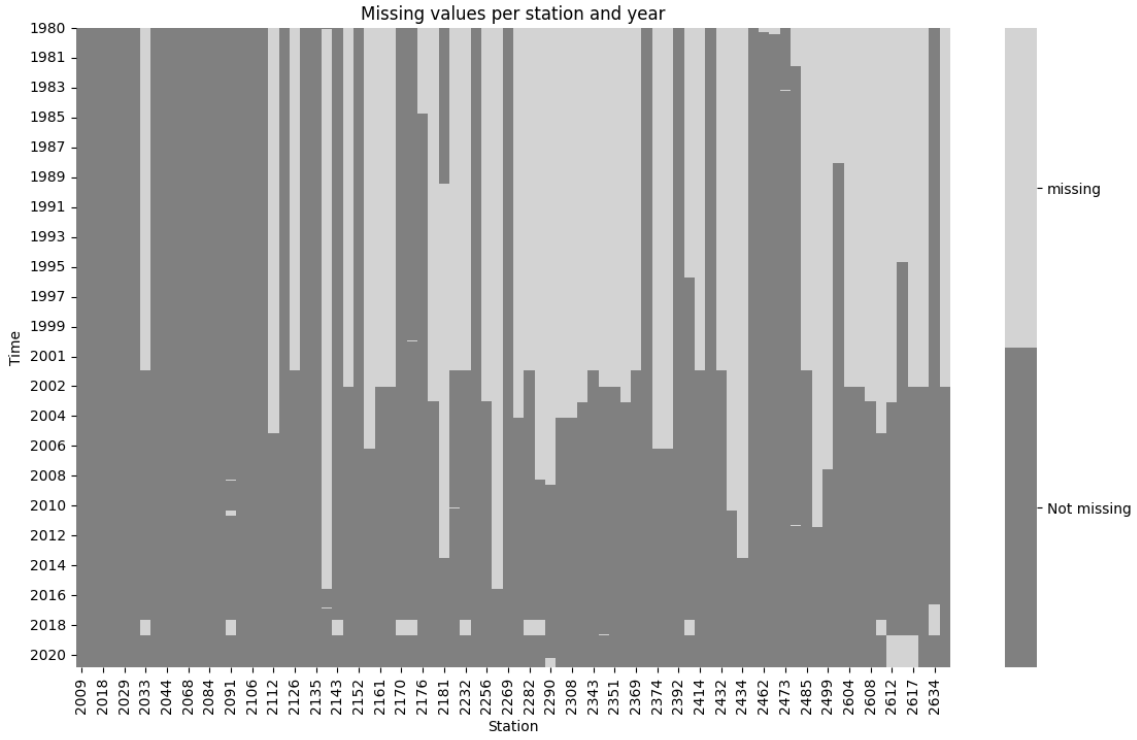


Figure 1.2: Heatmap showing missing values for each station

30% missing data while the 120 short gaps are barely visible. Another interesting point is, that in the year 2018 there are about 10 stations which did not record any data for exactly 365 days. In the air temperature and the discharge data are only 1% and 6% missing data respectively.

1.3.2 Challenges

The different LSTM models used, require different inputs. The best model available uses the air temperature, the discharge and the water temperature of neighbouring stations as input. Assuming all the data needed for the input for the LSTM models are present, there is no issue. But since there is about 30% of missing data in the water temperature, there are gaps that cannot be imputed with the LSTM models, since there may also be a gap in the input data needed for imputing another. Therefore do the different kind of gaps I categorized impose a challenge when imputing them. The short gaps pose no issue when it comes to imputing them, as the interpolation can be used, as already mentioned before, to approximate a value. As it can be seen in Fig. 2.2, interpolation delivers convincing results for short gaps despite its simplicity. The longer the gap gets, the less accurate the interpolation

will become. Therefore one of the LSTM models provided is applied, which also requires information on the air temperature, the discharge and/or the neighbouring station's water temperature. However the longer a gap gets, the more probable it is to encounter some kind of missing data in the input variables, namely the air temperature, the discharge or one of the neighbouring station's water temperature. In that case the gaps need to be treated individually to get the best achievable accuracy. I will go into more depth in Chapter 3. The three special cases mentioned earlier, which are also discussed in detail in Chapter 3 are the hardest kind of gap to impute, since many of their neighbours have missing data during a long period of time.

Chapter 2

Theory

In this chapter I will talk about previous research done in this field to put my work into perspective, give an overview of the different LSTM models I use and give a short introduction to the concept of LSTMs.

2.1 Previous Work

As already mentioned in the introduction, the accurate prediction of river water temperatures is an important factor for ensuring the health of an aquatic ecosystem. Therefore, several models have been developed over the years. in the following sections two models are introduced to put my methods into perspective.

2.1.1 Air2stream Model

The Air2stream model integrates a physically-based structure with stochastic calibration of its parameters [7]. This model serves as an intermediate approach between physically-based and purely statistical models. This hybrid approach retains the computational simplicity of statistical models while providing a physical basis for understanding river thermal dynamics. The model uses daily air temperature and river discharge as primary inputs to predict daily averaged river water temperature. To test the model's performance they applied this model to three rivers in Switzerland. The most complex model, which uses eight parameters achieves an RMSE of about 0.6°C during calibration and 0.72°C during validation on their dataset [7]. With these results, the model outperforms traditional regression models.

2.1.2 Gaussian Process Regression

GPR presents a probabilistic, nonparametric approach for solving nonlinear regression problems. Gibric et. al. use Gaussian Process Regression (GPR) to model both the long-term trends and the short-term variations in the river water temperature [8]. The framework hereby consists of two main components. The first component captures the seasonal patterns using a GPR model. The second component is used for the daily fluctuations caused by meteorological conditions using a second GPR model. The variables this second model uses are chosen based on their mutual information with the water temperature. They then combined those two GPR models to get their final model used for predicting river water temperatures. In order to validate their methods data from the Drava River in Croatia during the period from 1993 to 1998 was used. The best performing multiple GPR model achieved an RMSE of 0.87°C.

2.2 LSTM Models

The present work builds up on two recently published papers. In the first paper they introduce the graph structure that I am using for the information on the neighbours and the connection to the corresponding weather stations [6]. The graph structure of the river Rhine can be seen in Fig. 2.1. The blue dots indicate water stations that were built in the years of 1980 to 2000 and the red dots indicate water stations that were built after the year 2000. Annotated to each water station is their unique station number. In the second paper the different LSTM models for imputing water temperature are introduced [5]. These models are the basis of this work. In the following subsections I will introduce each model. Among these models are trained LSTMs which I will discuss later on in this chapter.

2.2.1 Interpolation

Interpolation is the simplest method to impute gaps. But as seen in Fig. 2.2, its performance starts to drop the longer the gap is and the other models outperform the interpolation. This limitation comes from the assumption of continuity in the data points during a gap, which becomes less valid for larger gaps, as the water temperature rises and falls during a year. As such, while interpolation remains a viable option for small gaps, more advanced techniques are necessary to handle longer gaps effectively.

2.2.2 A2Gap LSTM

This model is the first of the three LSTMs. It utilizes the air temperature to predict the water temperature. Due to the low performance in comparison to the more complex models, the A2Gap model will only be used where it is necessary. The more powerful models are utilized whenever possible.

2.2.3 AQ2Gap LSTM

AQ2Gap represents a refinement over A2Gap by incorporating additional information of the discharge (Q) of water stations. This model performs better than only using air temperature but relies on having recorded discharge values during the gap of the original station.

2.2.4 AQN2Gap LSTM

The AQN2Gap model uses additionally to the air temperature and the discharge the water temperature of neighbouring stations as a parameter. This model performs the best, particularly for longer gaps and it will be used whenever it is possible. While the AQN2Gap model has the lowest RMSE among all LSTMs, its effectiveness relies upon the availability and quality of the data. Nevertheless does its performance make it the preferred choice for imputing gaps.

2.2.5 LSTM Training and Performance

To evaluate and compare these models they chose 90 days sequences in the water temperature during which no gaps are present. They manually built in artificial gaps and used it as a training set. After imputing these gaps with the different models, they measured the root mean squared error (RMSE) of the predictions. The RMSE is defined as follows

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.1}$$

where n is the number of observations, y_i is the observed value and $\hat{y}_i$ is the predicted value.

The results of this evaluation can be seen in Fig. 2.2. As you may see, the interpolation is amongst the best models for the gap length of two days. For the gap length of 10 days the AQN2Gap model is the best, while the interpolation

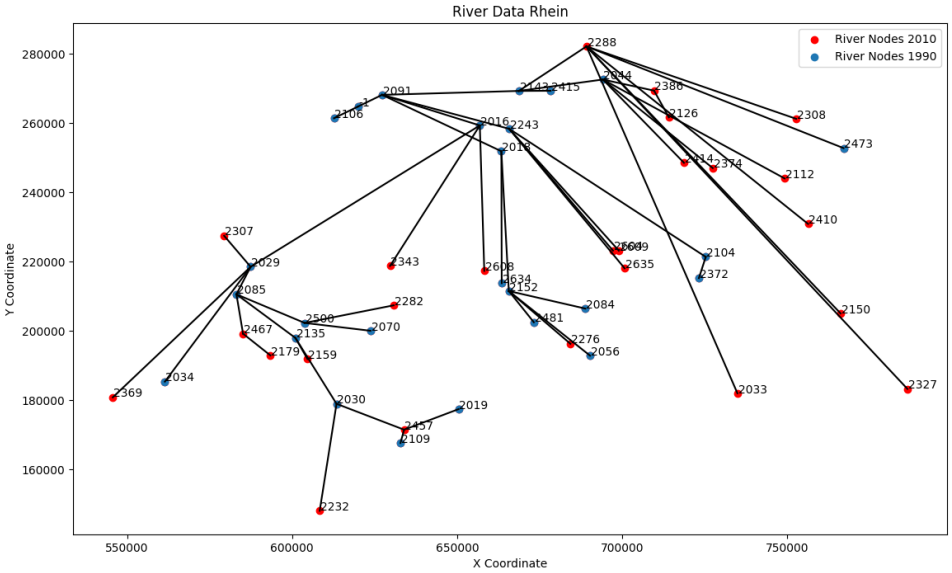


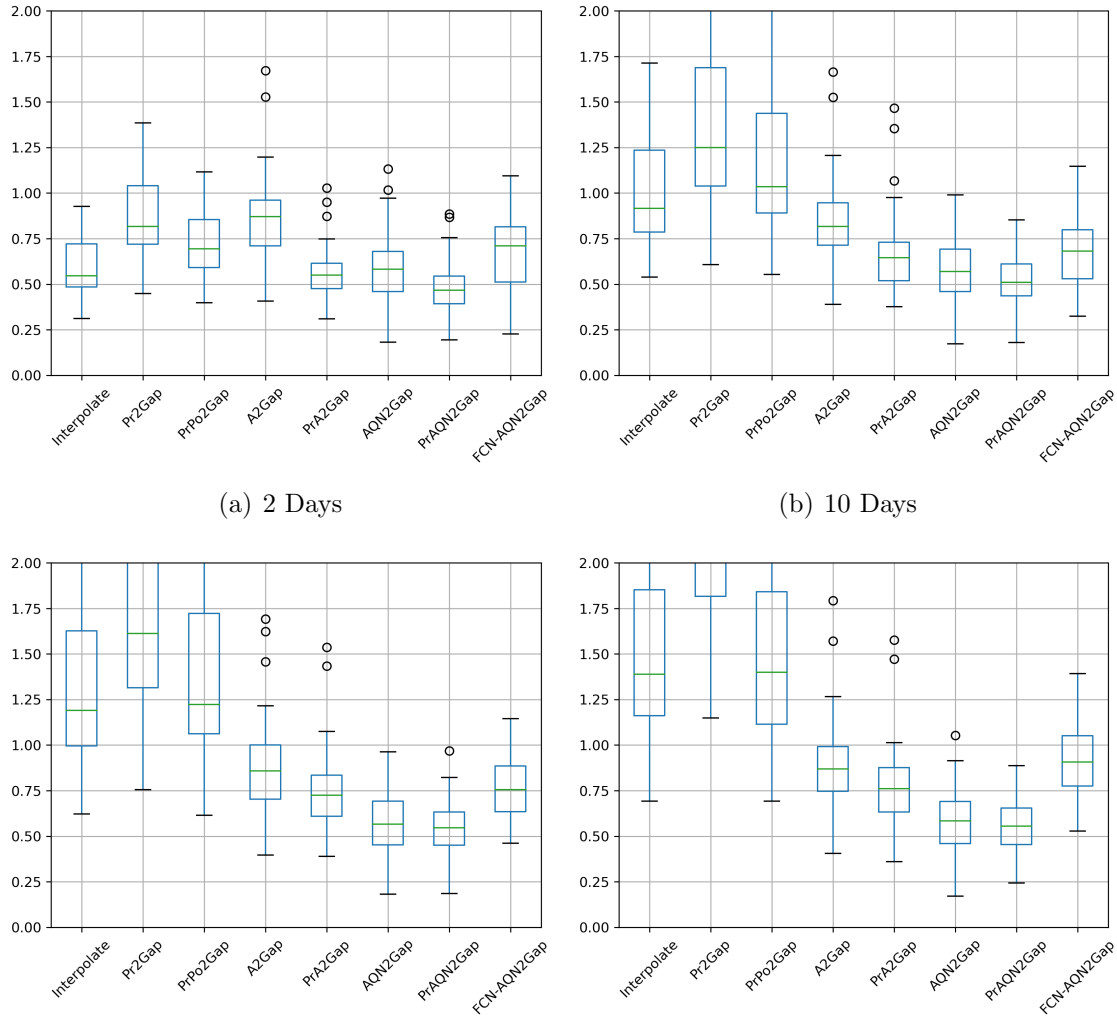
Figure 2.1: Rhine river network graph

is performing less good. This trend continues for the longer gap lengths. With these results in mind, I will try to use the better performing models, which use the neighbouring stations as a parameter, for the final prediction of the water temperature and use the A2Gap or AQ2Gap models only where it cannot be avoided due to missing data.

As already mentioned are these LSTMs trained by Fankhauser et. al. [5]. This means that for every station there are a few LSTMs ready to use with the corresponding information on the metadata like input size and the best weights. This has the advantage that the training was already performed and thus saving a lot of time. The downside of it is, that this limits the possibilities of dealing with missing data, as there is usually only one combination of neighbours for every AQN2Gap model.

2.3 LSTMs

LSTM, short for Long Short-Term Memory, is a special type of recurrent neural network (RNN) architecture designed to address the issue of vanishing or exploding gradients commonly encountered in traditional RNN's [9]. This problem arises due to the nature of sequential data processing, where information from previous time steps either decreases rapidly (vanishing gradients) or grows exponentially (exploding gradients) as it propagates through the network layers.



(a) 2 Days

(b) 10 Days

(c) 30 Days

(d) 60 Days

Figure 2.2: Results of the evaluation of different models [6]

LSTMs mitigate this problem by using a more complex memory mechanism. Unlike traditional RNN's, which only have a single recurrent hidden state, LSTMs feature multiple interacting memory cells that regulate the flow of information through the network. Each LSTM cell consists of three main components: the input gate, the forget gate, and the output gate.

The input gate controls the flow of new information into the cell, determining which information is relevant to retain. The forget gate decides what information to discard from the cell's long-term memory, preventing irrelevant or outdated information from persisting. Meanwhile, the output gate regulates the information flow from the cell's internal state to the rest of the network, ensuring that only relevant information is passed on to successive layers or time steps.

Furthermore, LSTMs maintain separate short-term and long-term memory states, allowing them to capture dependencies over long sequences more effectively. The long-term memory, also known as the cell state, serves as a stable storage unit that retains important information across multiple time steps without being subject to the same weight updates as the recurrent connections. This separation of short-term and long-term memory allows LSTMs to effectively capture and remember time-related dependencies in sequential data, making them particularly well-suited for tasks such as natural language processing, speech recognition, and time series prediction.

Chapter 3

Method for Water Temperature Imputation

This Chapter will discuss the strategy used to impute gaps of the Swiss river network in an accurate way using recent state-of-the-art deep learning models. Additionally will the strategies for the normal cases and the special cases be explained.

3.1 Data Preparation

Initially, we received the various datasets with the gaps and the corresponding LSTM models. In a first step, the encoding of the data is checked. In the air temperature datasets, missing values were represented by a "-", which is easily detectable.

The missing values in the air temperature and discharge datasets are not encoded in a special way, but are missing entirely from the data, i.e. the row for that specific date was not present in the dataset. In a first step a row was appended for every missing date with an added *NaN*-value as the temperature in the temperature column allowing the use of efficient python libraries such as pandas or numpy and to not run into a missing key error when wanting to look up the temperature on a day where it would be missing.

As the air temperatures do not come from the same stations as the water temperatures, a mapping had to be manually created between the two, indicating which water station is closest to which weather station. The mapping connects one water station, for instance 2009, to the three letter name of the water station, for instance *AIG*. As the files were all named *order_some-6-digit-number_data.txt* the files could not be read directly from the mapping. Therefore the contents of all the files are read into one big file which then could be subset with the three letter name.

Having a complete dataset with all the dates added in, allows to investigate the dataset more efficiently and work out the different types of gaps.

3.2 Gap Detection

In order to impute a gap, they first need to be identified. For this purpose the original dataset is examined and a new one gets created with the information for each gap on how long it is, the last day before the gap and the first day after the gap. For stations that were built in the years 2004 to 2010, the date 1980-01-01 gets noted as the starting date of the gap. With the these two dates the date range which is of interested can be created, in order to look up the air temperatures during this period. Information on the length of the gap makes it possible to differentiate the gaps that can be interpolated.

3.3 Model Preparation

As the LSTMs expect as input a vector with values between zero and one, they need to be normalized in a first step. The inputs are normalized using the minimum and maximum value of the specific variable, meaning that the minimum and maximum of the air temperature is used to normalize the air temperature inputs and so on.

After predicting the water temperature using one of the LSTMs, they will return a vector once again with values between zero and one. This vector needs to be denormalized using the corresponding minima and maxima, in this case using the water temperature. The formulas for the normalization 3.1 and denormalization 3.2 are

$$\hat{x} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{3.1}$$

$$x = \hat{x} \cdot (x_{\max} - x_{\min}) + x_{\min} \tag{3.2}$$

where x is the original sequence, $x_{\min}$ is the minimum value of the sequence, $x_{\max}$ is the maximum value of the sequence and $\hat{x}$ is the normalized sequence.

With the model preparations complete, the LSTM models can now be utilized to impute the gaps in the dataset. This process involves iterating over each station and applying each of the following models to the original data, saving the results of the individual models in different files.

3.3.1 Interpolation

As already mentioned, is the interpolation an effective solution for the short gaps. Because of this, interpolating the short gaps is implemented in every model, including the models for the special cases, hence no gaps with the length of 2 days or less will be imputed using one of these models, as they always get covered first.

3.3.2 A2Gap Predictions

In a first step the A2Gap model is used to impute every gap a station has. Using this method is fast and almost always applicable as only 1% of the air temperature data is missing [5]. This allows to impute almost every gap with the air temperature.

With all the datasets from the preparation in place, the air temperature values can be extracted with the help of the start and end dates. This results in a vector with the length of the gap that can be given as input to the corresponding LSTM after normalizing it.

The output received from the LSTM is again a vector the size of the gap length. the values can be assigned to the gap after denormalizing them.

Should a station have a gap in the water temperature as well as in the air temperature, the air temperature from a neighbouring weather station is looked up to be used instead to predict this gap. This should still yield good results, as the temperature should not differ by much between two neighbouring weather stations as the neighbours are usually close by. This also makes it possible to always use this model should the other models not be applicable.

3.3.3 AQ2Gap Predictions

The AQ2Gap LSTM is trained on the discharge additionally to the air temperature. Thus, in a second step the gaps in a station get imputed with the better performing AQ2Gap model. Hereby discharge measurements made during a gap in the water temperature get looked up and checked if any data, either in the air temperature or the discharge measurements, is missing during that period. The inputs to the LSTM are two vectors. One contains the air temperature and one contains the discharge measurements, once again both being normalized first. This method is

also applicable most of the time since only 6% of the discharge data is missing. Should it now be the case that a gap is present in the discharge data, these dates are ignored in my prediction and left empty, which leads to a gap in the resulting data. As seen in the evaluation, the AQ2Gap model achieves a lower RMSE than the A2Gap LSTM [5].

3.3.4 AQN2Gap Predictions

In a last step the AQN2Gap model is used to impute every gap. Once again will the discharge data and this time also the neighbouring stations get checked for missing data. This LSTM model expects additionally to the air temperature and the discharge measurements one vector for the water temperatures of each neighbouring station. To efficiently check if one of those values are missing, the discharge column from the respective dataset and the water temperature columns from each of the neighbours get appended to the dataset of the examined station. This way only one dataframe has to be checked for columns with missing data. The information on the neighbouring stations are received using the graph structure. An adjacency list is built using this structure, such that the neighbours a station has can be looked up easily. Using this adjacency list, every neighbour gets examined and the water temperatures get extracted from each dataset.

If data is missing for either discharge or water temperature from one of the neighbouring stations on a given day, it will be ignored and only dates for which all data is available will be imputed. This approach will once again lead to predictions with some gaps due to missing values in the input variables.

3.4 Strategy

For each of these models the predictions made with the respective model get saved in a separate CSV-File. After finishing with all the predictions, there are 4 files for every station. One file with only the original recorded data and one file for every model that a prediction was made with. The files with the prediction for every model may have gaps too due to missing data in the input variables. For the final evaluation of every station, the recorded data is checked for missing values. For every row with missing values in this data, the missing water temperatures get imputed with the ones from the AQN2Gap file, regardless if there are missing values in the AQN2Gap file. Additionally the string "AQN2Gap" is assigned to each row with a missing value, depicting which model was used to impute this gap. If there are still rows with missing values, they get imputed with the values from the next

best model and once again the model's name is assigned to each row. With this method it could be that the name with the model used for this prediction in some rows gets overwritten multiple times, as there may be missing values in more than one prediction. This method is continued until the file imputed with the A2Gap model is reached. After combining this information, the final dataset without any gaps and for each imputed value the information on which model was used is saved.

3.5 Limitations

One problem with the approach in the AQN2Gap model is, that regardless of how many neighbours are missing, this date in the dataset is ignored. This means, that if a station has three neighbours and only one of them has missing data, the data of the other two neighbours does not get used and instead one of the models which do not rely on neighbouring stations gets employed.

The reasons why in this case the A2Gap, AN2Gap or AQN2Gap model are not being used to predict a missing neighbour needed to get a prediction on a station are, that on the one hand, should there be an error in this prediction, this error will get propagated through the model into the currently examined station. There are models that can handle such a setting [5] but we provide a different solution to this limitation in the next chapter. On the other hand will a prediction on already predicted data generally perform worse than simply using either the AQ2Gap or A2Gap model. Let's say, the AQN2Gap model is used to predict a water temperature for the neighbour. As already seen in Fig. 2.2 this prediction has on average an RMSE of approximately 0.7. If we use this imputed data as a basis for a further prediction, this will lead to greater uncertainty, as the range of possible values tends to be larger, which is why it is more accurate to use one of the weaker models instead.

To still be able to make better predictions for the case where one or multiple neighbours are missing, three special cases are categorized and treated a differently.

3.6 Solving Special Cases

As already mentioned, there are a three special cases in the data for which the strategy does not perform well and which are more complex to solve. These cases mostly contain the situation where a new station was built several years later than its neighbours. Three cases are examined further which can be seen in Fig. 3.1.

On the left-hand side you may see the case where a new station (in red) was built between two already existing stations (in white). This situation is problematic

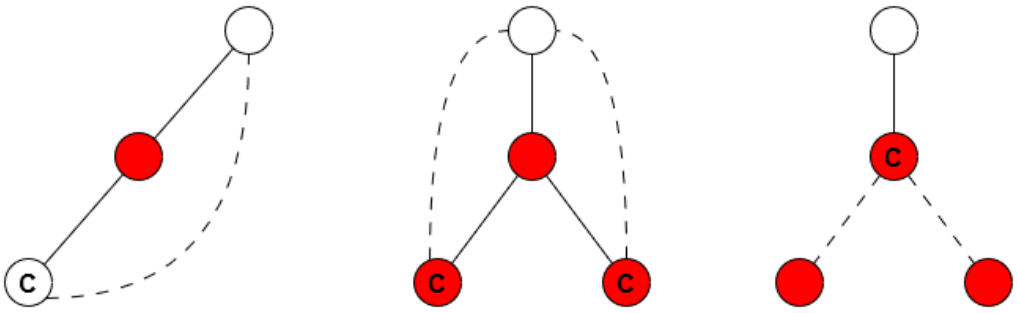


Figure 3.1: The three special cases

should one of the old stations have a gap during a time period where the new station was not yet built, therefore having no available data for the neighbours. To tackle this problem another connection is added, represented as the dashed line, to the two white stations, connecting them. So when the currently examined station, depicted with the **C**, is imputed, there is another neighbour available for use.

In the middle you may see the case where a new river branch was added. If the current station is either of the lower ones, once again marked with the **C**, either the A2Gap or AQ2Gap model must be used to impute this gap. This will result in a long period of time where the best model that is available (AQN2Gap) cannot be used. As seen in Fig. 2.2 the A2Gap model does not perform as well as the more advanced models, especially for longer gaps. To still be able to get good predictions and not having to resort to only the air temperature and the discharge to impute this gap, another neighbour is added to the stations, represented with the dashed lines. If a gap in one of the lower stations occurs during a time where the third red station did not record data yet, the alternative neighbour, depicted in white, is used as the only neighbour for the AQN2Gap LSTM. This means, that in this case the middle red station is ignored. Should on the other hand have the lower two stations missing data on a date where the middle red station was already built, only this original neighbour is considered for an estimation. This means, that in this case the connections to the white station are ignored.

The third special case is depicted on the right-hand side. The currently examined station has two neighbours each with missing values during the same period of time. Therefore can the standard AQN2Gap model not be used, as this will result in a gap in the prediction. But since the station has one neighbour with recorded data, the connections to the lower two red stations are discarded and only the white station is considered for the prediction. Should the lower two red stations have recorded data, will the LSTM trained on all three neighbours be used. For these three special cases new LSTM models were trained based on these changes in the graph structure. To have a better understanding of the graph with these special cases, a

new graph with only these stations and special connections is generated. One such graph is depicted in Fig. A.1 in the appendix. These three special cases can be seen in Fig. 2.1. The first one is for instance at the station 2457 and the second and third one can be observed on the right hand side at the station 2288 with most its neighbours. Because these special cases need special connections where either new neighbours were introduced to a station or existing neighbours were omitted, an adjacency list with only those special connections is created. Additionally a special function specifically designed for those special cases was coded, where the neighbours get looked up using this special adjacency list and the special LSTM models trained on those special connections get loaded. Should a station not be a key in this adjacency list it can be skipped over, as it does not need to be covered.

Using this strategy allows using the AQN2Gap model even more which will lead to an overall better prediction of the real world data, as this model has the lowest RMSE out of the models that get used.

One challenge of using trained LSTMs is the amount of models that need to be stored and loaded. When imputing the water temperature for almost 60 stations and each station has three to four LSTM models, depending on whether it has a special case that gets considered, will add up to more than 180 different models that need to be managed. In order be able to automate loading the specific model, a consistent naming convention is introduced as the models get loaded using the file names. The naming convention followed the scheme of starting with *station_model*, as example *2009_aqn2gap*. This approach ensures that models can be efficiently identified and retrieved based on their filenames.

Furthermore, each of these models generates predictions spanning over 40 years of data, making a verification process extremely time-consuming. Making sure that each model's predictions are accurate and function as intended requires careful examination and validation which is hardly doable with this amount of data.[1]

[1]The code is available under: `https://github.com/CarloRobbiani/swiss_rivers`

Chapter 4

The Gap Free Swiss River Network

This chapter discusses the results achieved with the strategy. It will compare the results of each different strategy and how the effectiveness of them is measured.

4.1 Measuring Results

Direct comparison of my results to a ground truth is not possible because the true value of the missing data points are not known. In order to measure how good my strategy works, the frequency of different model usages is used as an indicator for the performance. When looking again at the results of the evaluation in Fig. 2.2, the usage of the AQN2Gap model should be maximized in order to get most accurate results. Using the AQN2Gap for imputing the gaps yields better results than using only the A2Gap or AQ2Gap models. The results can be seen in Table 4.1 but I will go into more detail in the following section. These percentages were calculated based on the column where the model type is stored, after filtering out every row that has observed data.

4.2 Qualitative Analysis

In this section the results of the strategy are discussed and some examples are shown. In Fig. 4.1 you may see the imputation results of the different models for one station. The gap shown in the figure has a length of 365 days spanning from the 01.01.2018 to the 31.12.2018. The blue line hereby indicates recorded data and the red, orange, green and black line segments each indicate data that was imputed with the respective model. The interpolation, represented with the black

line, struggles to correctly represent the trends in the water temperature. That is why it should not be used for long gaps. As seen, even when using only the air temperature as an input, the imputed data aligns well with the other models and the trends in the water temperature, showing that even the simplest LSTM out of my models can effectively capture the underlying patterns and trends. This initial A2Gap model sets a solid baseline for imputation accuracy. The results from the other models, seen in the orange and green lines show a similar result. As you may observe did every model predict a value for this specific gap. Hence will the final dataset for this particular station be imputed with the temperatures from the AQN2Gap model, since this is the best model available to me. Note however, that if there had been missing data in the neighbouring stations, there would be no prediction for the AQN2Gap model at certain dates. This would lead to a gap in the respective prediction file, i.e. the AQN2Gap file and the AQ2Gap model needs to be used for the final prediction. This is the major limitation of the AQN2Gap model.

In Fig. 4.2 you may see a zoom in of Fig. 4.1, which better illustrates the differences in the predictions of the various models. While in Fig. 4.1 the predictions may seem close to each other most of the time, you may observe in this figure, that the predictions of the different models show a difference of up to 2 degrees Celsius for a given day, with the biggest gap between the AQ2Gap and the AQN2Gap models. These differences show nicely the impact of different input variables on the imputation results. Additionally, the graph highlights the importance of model selection in achieving accurate temperature predictions, as even models with a similar architecture can yield significantly different outcomes. This graph also demonstrates the difficulty when having to choose the model for the final prediction. Without any ground truth to compare the predictions with, I have to trust the evaluation results in Fig. 2.2.

To better visualize this problem and to compare the performance of the different models, I manually built in an artificial gap into one of the stations with a length of about 6 months. This gap was then imputed with the different LSTM models. In Fig. 4.3 the results of the different models are depicted, along with the real world data. Note that this Figure only shows a section of the gap to better visualize the differences. As you may see, are the models most of the time closely aligned with each other and the observed values. There are still some dates during which no model manages to predict the temperature correctly and even the AQN2Gap model shows an error of about 0.5°C. The A2Gap model seems to have the most trouble with predicting the real world data, as it shows the biggest distance to the blue line. This fits in well with the results of the evaluation which states, that the

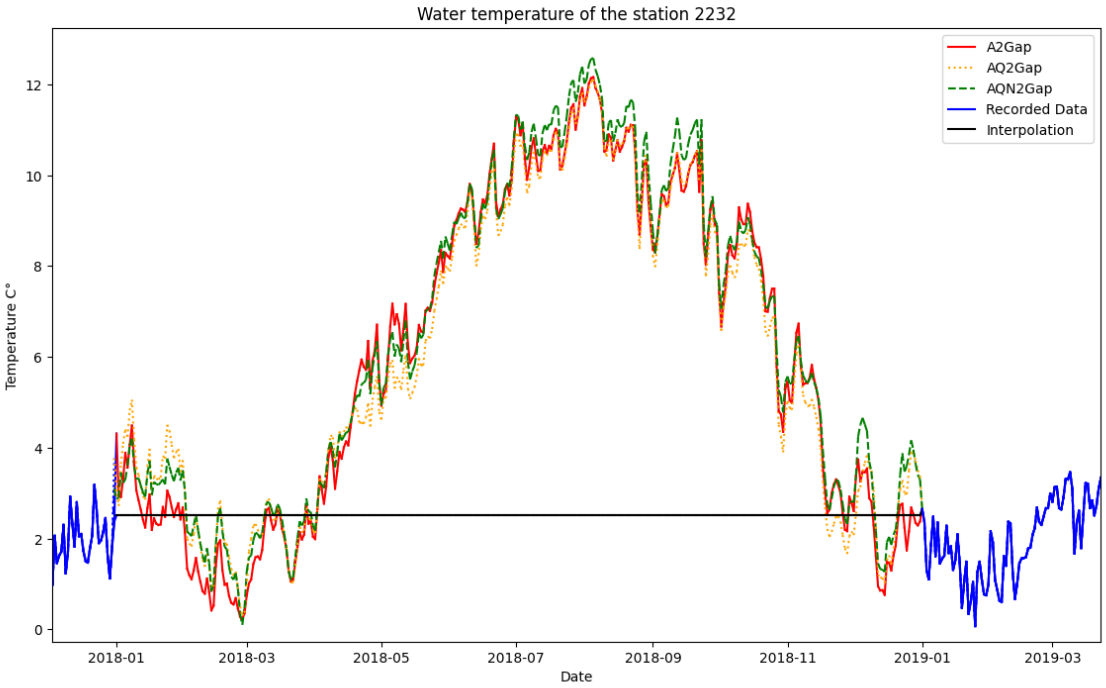


Figure 4.1: Results of the different models

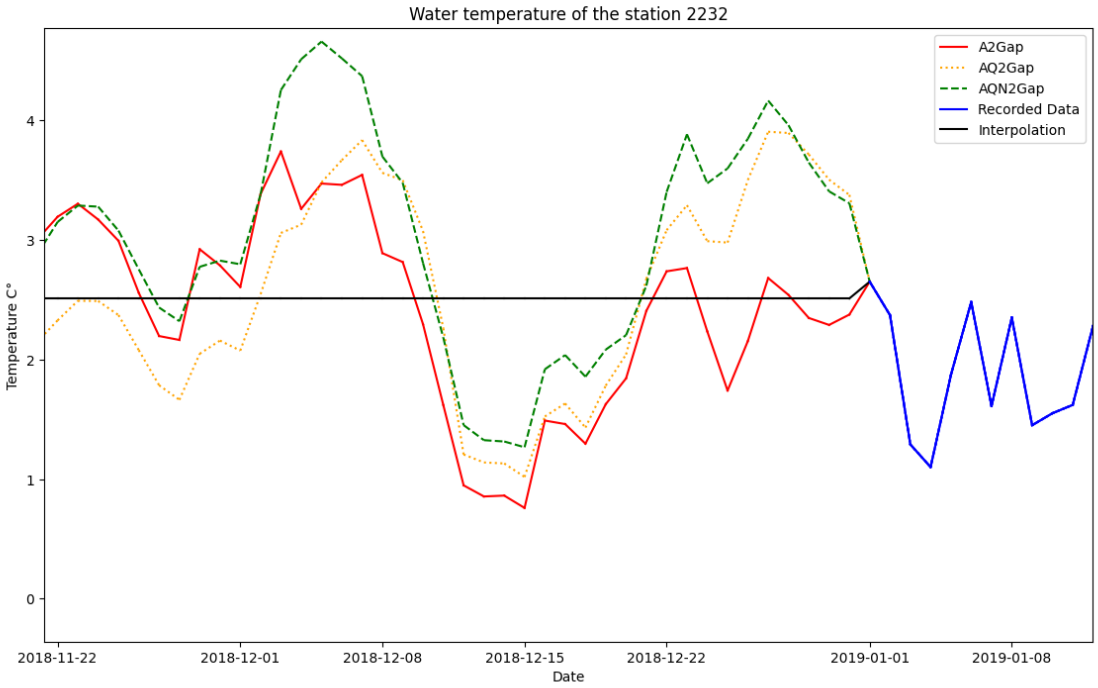


Figure 4.2: Results of the different models

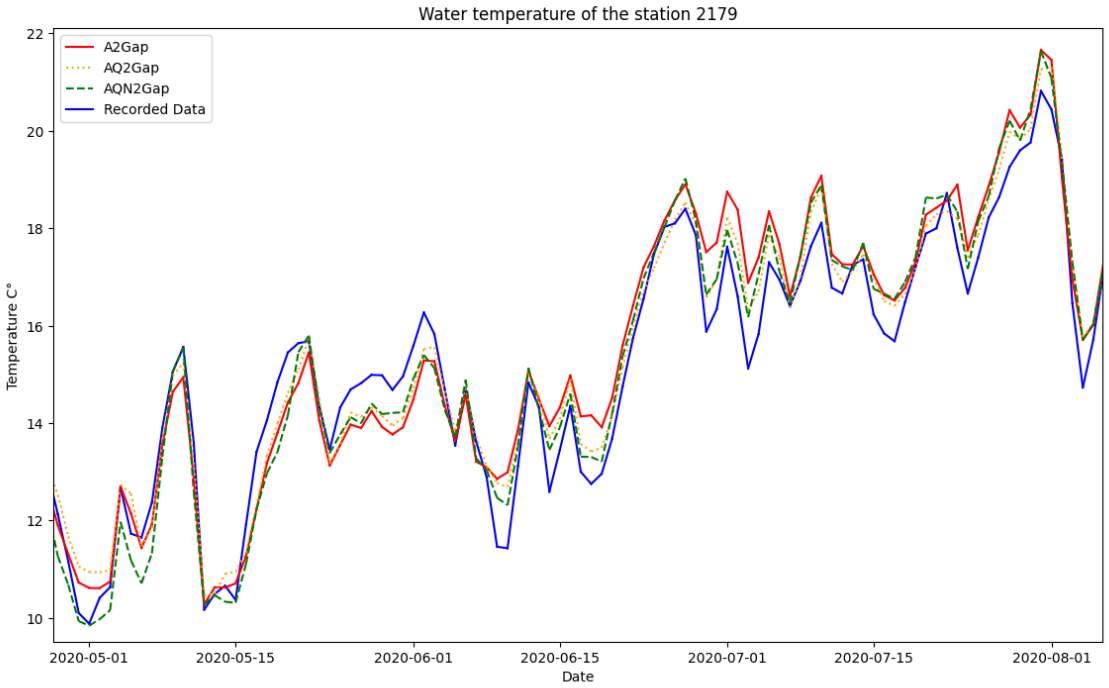


Figure 4.3: Comparison of models with observed data

A2Gap model performs the worst out of the models used in this paper, especially for the longer gaps. The AQ2Gap and the AQN2Gap models on the other hand do not show a significant difference most of the time. Although there are some dates during which the AQ2Gap model delivers slightly more accurate results than the AQN2Gap model, the latter is chosen for the final prediction as it overall achieves better results.

4.3 Quantitative Analysis

In the heatmap in Fig 4.4 , the distribution of the gaps for each station is depicted. This time the gaps are colored in depending on which model was used to impute the gap. Note that since the interpolation is used in every model, the gaps imputed with the AQN2Gap model also include the ones imputed with interpolation. I decided not to make a separation between these two, as the short gaps are barely visible on the heatmap and the interpolations performance for such gaps is comparable with the AQN2Gap model. The results shown in this heatmap do not include any special cases yet, meaning that if any data is missing, the respective model did not get used. The final estimation rarely uses the A2Gap model for the water temperature, while the AQ2Gap and AQN2Gap models seem to be the most dominant for the final

prediction. The A2Gap model is only used for stations who have longer sequences of missing data in the discharge, which is the case for roughly ten percent of the stations. As you may notice, are there fewer stations depicted in this heatmap than in the heatmap in Fig. 1.2. The reason for that is, that there are about 20 stations which are not connected to any other water stations and therefore are not depicted in the river network graph. As I did not have LSTM models for these stations I did not cover them in the predictions, but since those do not possess any neighbours the AQ2Gap model would be used to impute the gaps in them.

In the table 4.1 is the percentage each model was used for the final prediction in this first approach. The A2Gap model was not used frequently with only 9.83%, as it only is used if the discharge is missing which lines up with the 6% of discharge data that is overall missing in the dataset. The AQ2Gap model was used in 38% of the cases. The AQN2Gap model is the most used one with 52.17%. While these results already look promising, there are still some water stations with longer periods of time where the AQN2Gap model could not be used. This is mostly the case because some or all neighbours of those stations also have missing water temperatures. To tackle this problem, I incorporated the strategy for the special cases to further improve these first results. The goal of these special cases is, to use the AQN2Gap model even more frequently and hence improve the accuracy of the overall prediction.

To be able to use the AQN2Gap model more frequently, different strategies were implemented as already mentioned in Chapter 3, which use LSTMs that are trained on special connections. These special connections mostly involve newly built station, since it is the missing neighbours which prevent the AQN2Gap model from being used. When employing this strategy, it should allow the AQN2Gap model to be used more often.

As you may see in the table 4.1, using this method yields better results since the AQN2Gap model gets used more often. When looking at the heatmap in Fig. 4.5 you may also see that the AQN2Gap model is used more frequently but there would still be room for improvement. Especially the gaps which occur in the year 2018 did not change at all using this strategy. This is the case, because the gaps during this year show a different constellation of missing neighbours than the gaps earlier. Therefore the special LSTMs trained on the missing neighbours for the longer gaps could not be used, as the special connections would need to be setup differently. In order to cover these gaps too, further LSTMs would have to be trained which also cover this configuration of missing neighbours. The last cases that were covered are the ones where there are only some neighbours missing, allowing the AQN2Gap model to be used even more. As you may notice, using these special cases did

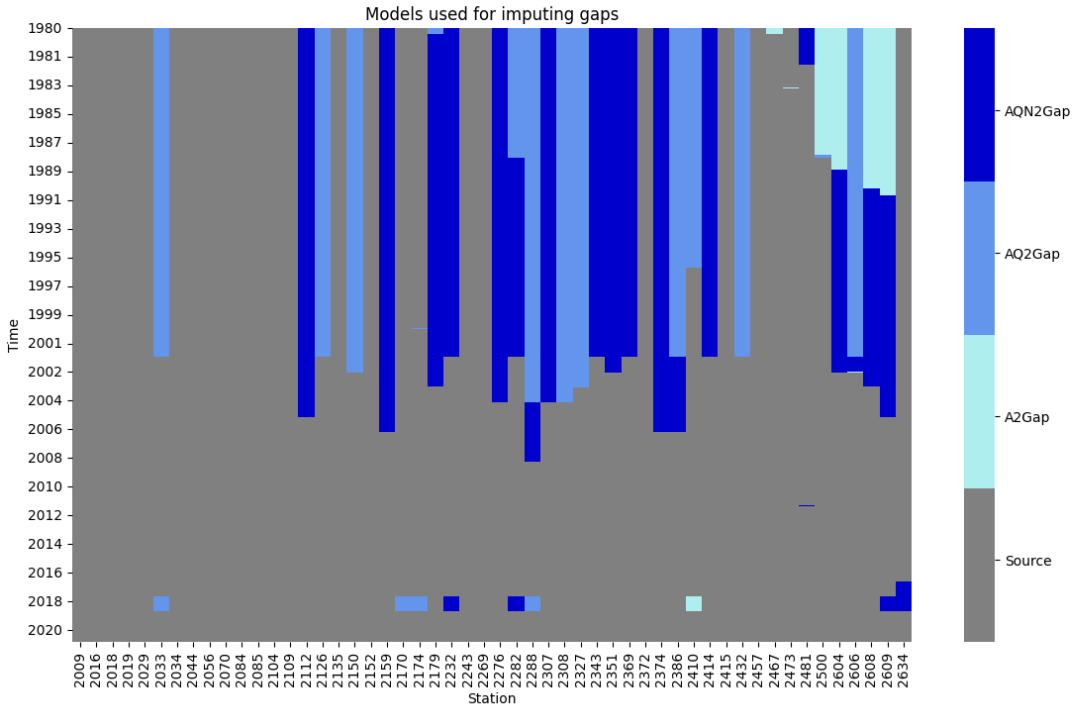


Figure 4.4: Heatmap with different models

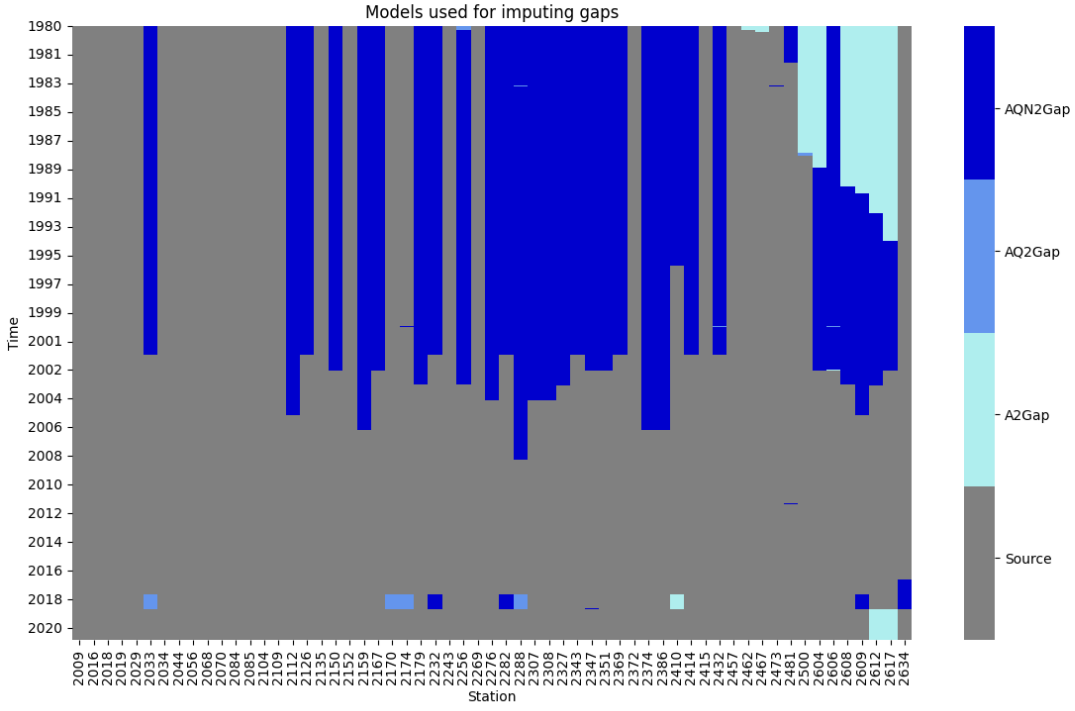


Figure 4.5: Heatmap with different models with incorporated special cases

Model	No special cases	With special cases	Difference
A2Gap	9.83%	9.83%	±0
AQ2Gap	38.00%	0.62%	−37.27%
AQN2Gap	52.17%	**89.56**%	**+37.27**%
RMSE	0.732	**0.652**	**-0.08**

Table 4.1: Percentage of times each model was used

not decrease the share of the A2Gap model, since they do not cover the case of missing discharge. Dealing with the missing discharge is harder than dealing with missing water temperatures, since the discharge often varies significantly between two neighbouring stations. After incorporating all the special cases into the results, I achieved a coverage of the AQN2Gap model of 89.56%, with only 0.62% of the time where the weaker AQ2Gap model is used.

These results show the effectiveness of the different strategies. Using the more advanced ones enables the AQN2Gap model to be used for 37.27% more gaps when compared to the standard strategy of not using a model when inputs are missing.

Every LSTM model has information on the RMSE it achieved. Utilizing this RMSE from each trained LSTM model, the weighted RMSE of the final prediction is computed. In the absence of any special case considerations, an RMSE of 0.732 is obtained. This RMSE is calculated by adding up the model RMSE of each prediction and then dividing by the total number of predictions made. When incorporating the special cases, the RMSE further improves to 0.652, marking an enhancement of 0.08.

Chapter 5

Conclusions and Future Work

In this chapter I will revisit and discuss the central questions of the paper.

5.1 Conclusion

The long-term data on stream temperatures is essential for understanding the thermal regimes of riverine ecosystems [4]. River temperature is crucial for determining the health of aquatic ecosystems, as each species has a specific temperature range it can tolerate. Significant changes in river temperatures can adversely affect these species. Therefore, collecting data on river temperatures is vital for comprehending long-term changes, particularly in the context of climate change.

The Federal Office of the Environment (FOEN) in Switzerland has been collecting water temperature and discharge data from Swiss rivers since 1980. However, gaps in this water temperature data, often caused by malfunctioning sensors, planned maintenance, or communication issues, can impact data analysis, leading to potentially inaccurate results. Additionally, I had access to atmospheric data, specifically air temperature measurements from MeteoSwiss.

The goal of the present work is to use three existing LSTM models that were trained on the discharge, air temperature measurements and water temperatures of neighbouring stations in order to predict water temperatures in various water stations placed in the four big rivers Rhine, Rhone, Inn and Ticino. I successfully imputed all gaps in my dataset in a first approach with one of the available models. With the first approach the AQN2Gap model could be used of for 52% of the cases and a the AQ2Gap model for 38% of the cases when imputing every gap. To further improve on these results, I identified and examined three special cases. These special cases mostly involve the situation where a new station was built in between two already existing stations or a station either has some missing neighbours or only missing neighbours for a long period of time. In order to tackle these problems,

new neighbours were added or existing neighbours were ignored and based on those transformations in the graph structure, new LSTMs were trained to accommodate those changes. With this special approach for certain stations, satisfying results are achieved where the best model available, namely the AQN2Gap model, could be used for over 80% of the gaps.

5.2 Future Work

In order to make my predictions I had access to trained LSTMs. These LSTMs were not trained on every possible combination of neighbours, meaning that for a station with four neighbours there was only one LSTM trained for the AQN2Gap model. This model is expecting additionally to the air temperature and the discharge a value for every neighbour it has, i.e. six values in this case. This leads to the situation that if one neighbour is missing on a given date, this model cannot be used for a prediction since it is lacking one input. To further improve my results, more different LSTMs could be trained for more combinations of neighbours. With more models trained on different amounts of neighbours for every station, the AQN2Gap model could be used more frequently which will lead to a more accurate prediction overall. Another approach could be to take more input parameters, for instance the humidity or the depth of the river, to further improve the predictive capabilities of the LSTM models.

In this paper the three most prominent special cases were covered. One could investigate other special cases further and try to improve the final predictions by implementing different strategies for them. Another special case one could look into is, that there are about 20 water stations which are not connected to one of the big rivers Rhine, Rhone, Inn or Ticino. These stations therefore have no available neighbours and you would need to take either the A2Gap or AQ2Gap model for an estimation. To be able to use the better models with the neighbours as a parameter, you could either connect it to the nearest neighbouring station or if that is not possible find some other way to get more parameters.

As seen on the heatmap in Fig. 4.5 there are about 9% of the gaps which have missing values in the discharge measurements. As already mentioned are these gaps difficult to tackle, as the discharge varies a lot between two neighbouring water stations. One could come up with different strategies to get values for the discharge during this time. One possibility could be to train some Machine Learning model on the discharge with the goal of imputing the gaps in the discharge datasets first. Another approach could be to take the discharge values of another water station that has the lowest deviation in the discharge.

Additionally instead of restricting the work on the water temperature, one could also investigate the gaps in the discharge or the air temperature and try to impute them with some Machine Learning techniques. Having a more complete record on these measurements would in turn also help with the predictions of the water temperatures.

Appendix A

Appendix Title

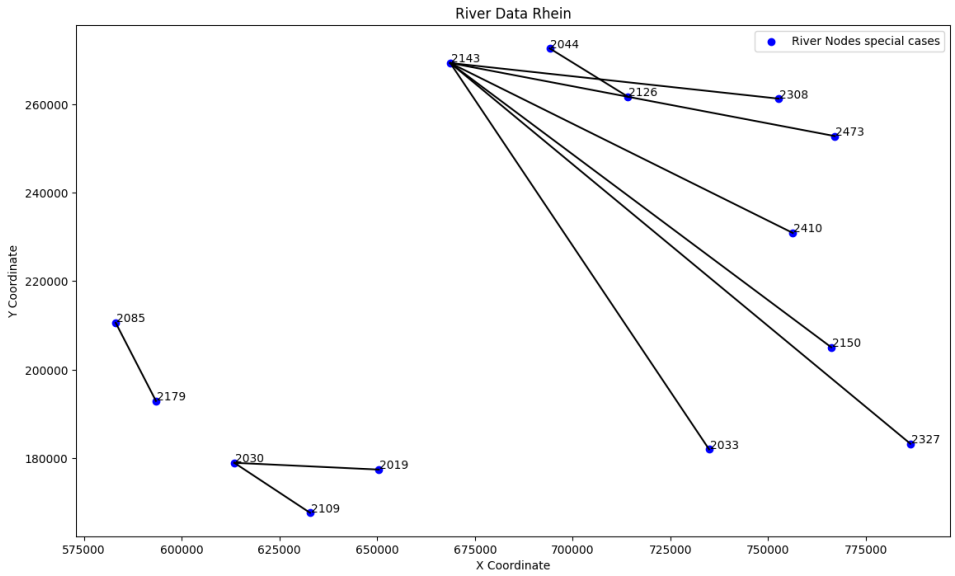


Figure A.1: Special graph showing the special connections for the river Rhine

Bibliography

[1] Daniel Caissie, Mysore G. Satish, and Nassir El-Jabi. Predicting water temperatures using a deterministic model: Application on miramichi river catchments (new brunswick, canada). *Journal of Hydrology*, 336(3):303–315, 2007.

[2] Florentina Moatar and Joël Gailhard. Water temperature behaviour in the River Loire since 1976 and 1881. *Comptes Rendus. Géoscience*, 338(5):319–328, 2006.

[3] Paulin Hardenbicker, Carsten Viergutz, Annette Becker, Volker Kirchesch, Enno Nilson, and Helmut Fischer. Water temperature increases in the river Rhine in response to climate change. *Regional Environmental Change*, 17(1):299–308, January 2017.

[4] Bruce W. Webb, David M. Hannah, R. Dan Moore, Lee E. Brown, and Franz Nobilis. Recent advances in stream and river temperature research. *Hydrological Processes*, 22(7):902–918, 2008.

[5] Benjamin Fankhauser, Vidushi Bigler, and Kaspar Riesen. Graph-based deep learning on the swiss river network. 14121:172–181, 2023.

[6] Benjamin Fankhauser, Vidushi Bigler, and Kaspar Riesen. Impute water temperature in the swiss river network using lstms. pages 732–738, 2024.

[7] Marco Toffolon and Sebastiano Piccolroaz. A hybrid model for river water temperature as a function of air temperature and discharge. *Environmental Research Letters*, 10(11):114011, nov 2015.

[8] Ratko Grbić, Dino Kurtagić, and Dražen Slišković. Stream water temperature prediction based on gaussian process regression. *Expert Systems with Applications*, 40(18):7407–7414, 2013.

[9] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–80, 12 1997.