ECMI 2023 - Hydro

1 Overall strategy

1. Reduce dimension of problem
2. Choose useful data
3. Choose a model
4. Train a model
5. Evaluate results

2 Choosing stations

We have water level data from 56 stations, new/old, close/far, up-/downstream. Reducing the number of stations considered will help us train our model more efficiently. For example, we would like to only consider staitions that are close enough to influence the water level in szeged within 7 days.

To do this we applied several statistical approaches:

2.1 Shifted correlation

For $i \in I = \{0, \ldots, 30\}$

Calculate the correlation $c_{i,X}$ of the water level between Szeged at day d and Station X at day $d - i$.

$t_X = \text{argmax}_{i \in I} c_{i,X}$ then gives an estimate on how long it takes for X to influence Szeged.

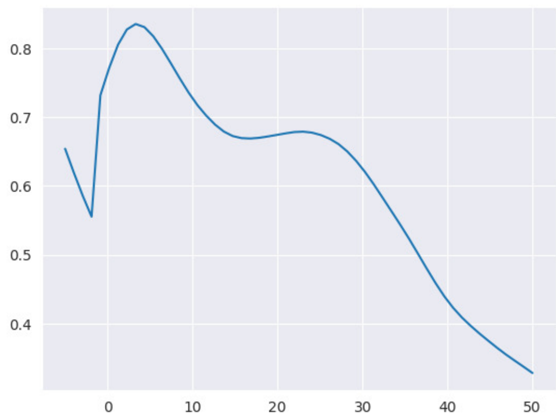

Figure 1. Graph: x-axis: i, y-axis: c_i for a Station near _, $p_{\text{val}}(t_X) \ll 10^{-3}$, 2004-2006

If we now visualizes t_X for every station on a map we can see if the results atleast appear to be making sense.

TODO

Figure 2. Stations colorized by t_X

2.2 Different types of correlation

In the procedure above we initially used the regular Pearson correlation. We then repeated the procedure with Spearman and Distance correlation. We got similar results:

TODO

Figure 3. Map, Spearman Correlation

TODO

Figure 4. Map, Distance Correlation

2.3 Causatily

As Correlation does not necessarily mean causality we also used used granger causality:

TODO

Figure 5. Map, Ganger Causality

TODO choose stations for prediction

3 Selecting Training data

As the riverbed constantly changes (errosion, floods, dams, ...) we want to choose the training data carefully. If we would for example do the usual 80/20 spilt of training and validation data and then train on the older 80% and validate on the newer 20% then the model mostly knows how to predict floods 50 years ago but maybe not modern floods as the river system has changed to much. So we want to spilt the data in a more soffisticated way. In order to train and validate on parts the newest data.

To get an idea on how to spilt the data we did research on the floods and dams of the Tisza

Name	since	lon.	lat.	Wattage
Gibárti vízerőmű	1903	48.317944	21.1635	1000
Felsődobszai vízerőmű	1911	48.263687	21.084688	940
Békésszentandrási duzzasztó	1942	46.891	20.4995	2000
Kesznyéteni vízerőmű	1945	47.99597	21.033205	4400
Tiszalöki vízerőmű	1959	48.025141	21.307876	12900
Kiskörei vízerőmű	1973	47.492961	20.515569	28000

Table 1. general information about the dams

	from	to	height in cm
1970	May	June	961
2006	April	May	1009

Table 2. biggest floods in Szeged

TODO: map

TODO: does this influence our choice in stations? compare with chapter 2.

4 Choosing a model

We are considering 3 different models.

i. The **LSTM** (= Long-Short Term Memory) takes the water level data of a single day as well as its memory as an input and outputs a prediction for the next day as well as its new memory for the next day. To get a good prediction one has to feed the model data from several consecutive days leading up to the present.

ii. The **TFT** (= Temporal Fusion Transformer) takes water level data from many days as input and will output a prediction. This prediction could either be for all stations on the next day (recurrsively predict 2 to 7 days ahead) or just the 7 day prediction for Szeged.

iii. A NN that takes a **fixed number of days** as an input and provides a prediction (again either a single day prediction for all stations or the 7 day prediction for Szeged directly).

    This approach does not yet incooperate the ordering of the inputs into account. Which is generally not the best approach as the model will have to figure it out during training.

    One could use a **CNN** (= Convolutional Neural Network) which convolutes over the time dimension of the input. The structure of such a CNN would suggest the recurrsive one day prediction strategy.

TODO