PREDICT_GROUNDWATER

A software prototype for predicting groundwater fluctuations in past and future climates using machine learning.

By Philip Groß, April 2024.

Disclaimer: This software and the presented results were developed over the course of only two weeks. The material is therefore of very preliminary nature and MUST NOT be taken as policy advice. It serves solely as a case study for potential further research and development.

A: Project description

In many parts of the world, groundwater is needed as a source of drinking water for humans or used in agricultural and industrial activities. Therefore, accessible groundwater is a vital resource for many people. However, groundwater levels do not remain constant, but are subject to perpetual change, which reflects the variable nature of the many parameters that ultimately control groundwater levels. These are mainly meteoric (e.g. precipitation, temperature, vegetation) or geologic (soil and rock properties) factors, but also human activity can directly influence groundwater levels (e.g. groundwater extraction via wells).

The broader goal of this project is to investigate the influence of meteoric conditions on groundwater levels over a long timespan. In the first part, this is performed for the past of a selected region. A machine learning model is trained on meteoric data to predict the observed groundwater movements of the past. The second part deals with using the trained model to predict the groundwater behavior in hypothetical future climate scenarios.

A1: File inventory

The repository is organized in the following way:
- folders containing weather, groundwater and elevation data for the model region
- notebooks performing special tasks, organized in a modular way
- a file called toolbox.py containing custom helper functions for use across notebooks

A2: Modeling strategy

The modeling strategy involves consecutively running the notebooks to perform the required modeling steps. Each notebook generates output that is then used by successive ones.
1. Data retrieval using get_weatherdata and get_gwdata_bergstraße.
2. Model generation and export using model_1.
3. Creation of future weather scenarios using model_weather.
4. Predict groundwater with the trained model and a weather file using future_model.
5. Analyze modeling results using analyse_predictions and make_animation.

B: Results for Bergstraße model region

This section shows the modeling process step-by-step for the Bergstraße region, which is located in the eastern Upper Rhine plain (Germany) between Mannheim/Heidelberg in the south and Darmstadt in the north.

B1: Data retrieval

Start with meteorological data using get_weatherdata.ipynb. In the beginning of the notebook, specify the region from which data is to be downloaded from the DWD database by entering x- and y-coordinates in decimal degrees. The weather stations used in this study are shown in Fig. 1. After specifying the coordinates, the notebook will perform the following operations:

- Download metadata for all stations in the region.
- Download daily weather data for each station over the entire available timespan.
- Some data cleaning.
- Calculation of derived weather data, e.g. daily averages for the entire region.
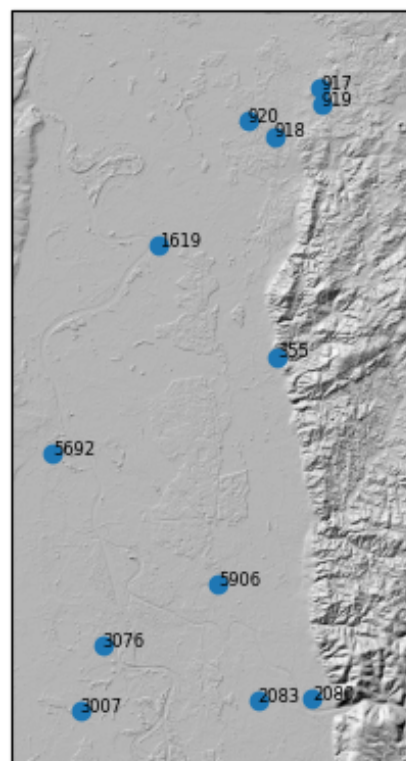- Storing the data in external files for later use.

Fig. 1. Location of the used weather stations in the upper Rhine plain between Mannheim/Heidelberg (south) and Darmstadt (north).

Groundwater data: Groundwater stations in Germany are operated by the individual states, and no general data retrieval method is available. Since the model region is located in the state Hesse, we need to use their download portal to get the station data manually. Fig. 2

shows the dense spatial distribution of groundwater stations used here. The notebook get_gwdata_bergstraße.ipynb (see there for details) then processes this data further:

- Loading and cleaning of the station meta- and measurement data for the entire available timespan.
- Some consistency improvements and reordering.
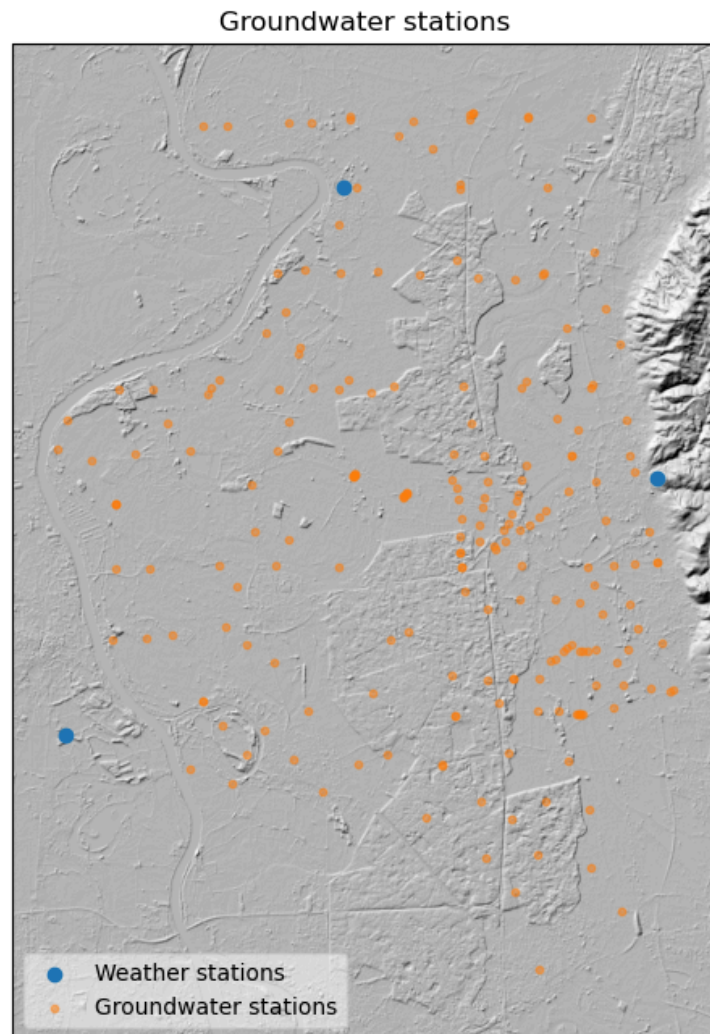- Concatenating and saving the data in external files.

Fig. 2. Location of the used groundwater stations in southernmost Hesse in the center of Fig. 1, roughly between Mannheim/Viernheim (south) and Pfungstadt (north). The main rivers are the Rhine, running along the western map margin and the Weschnitz, running southeast-northwest across the map and into the Rhine.

B2: Model training

The model training is performed in the model_1.ipynb notebook (or ones with similar names). It performs the following steps:

- Loading of all necessary data tables to data frames.
- Creation of station clusters for location encoding.
- Data merging and restricting the timespan to 1950-2022.

- Train-test split on a station level, i.e. stations (with their entire data timespan) are randomly assigned to either the train or test set.
- Cascading (i.e. exploratory and refinement stages) hyperparameter search strategy. This is where the regression algorithms and their parameters for testing are selected. Note that a grid search is very computation-intensive for most algorithms. Note also that the current cross-validation strategy does not split the data by stations but by dates, which is potentially problematic and could be improved in the future by using a custom-made cross-validation.
- Inspection of model performance on the test data (see Figs. 3 and 4 for examples).
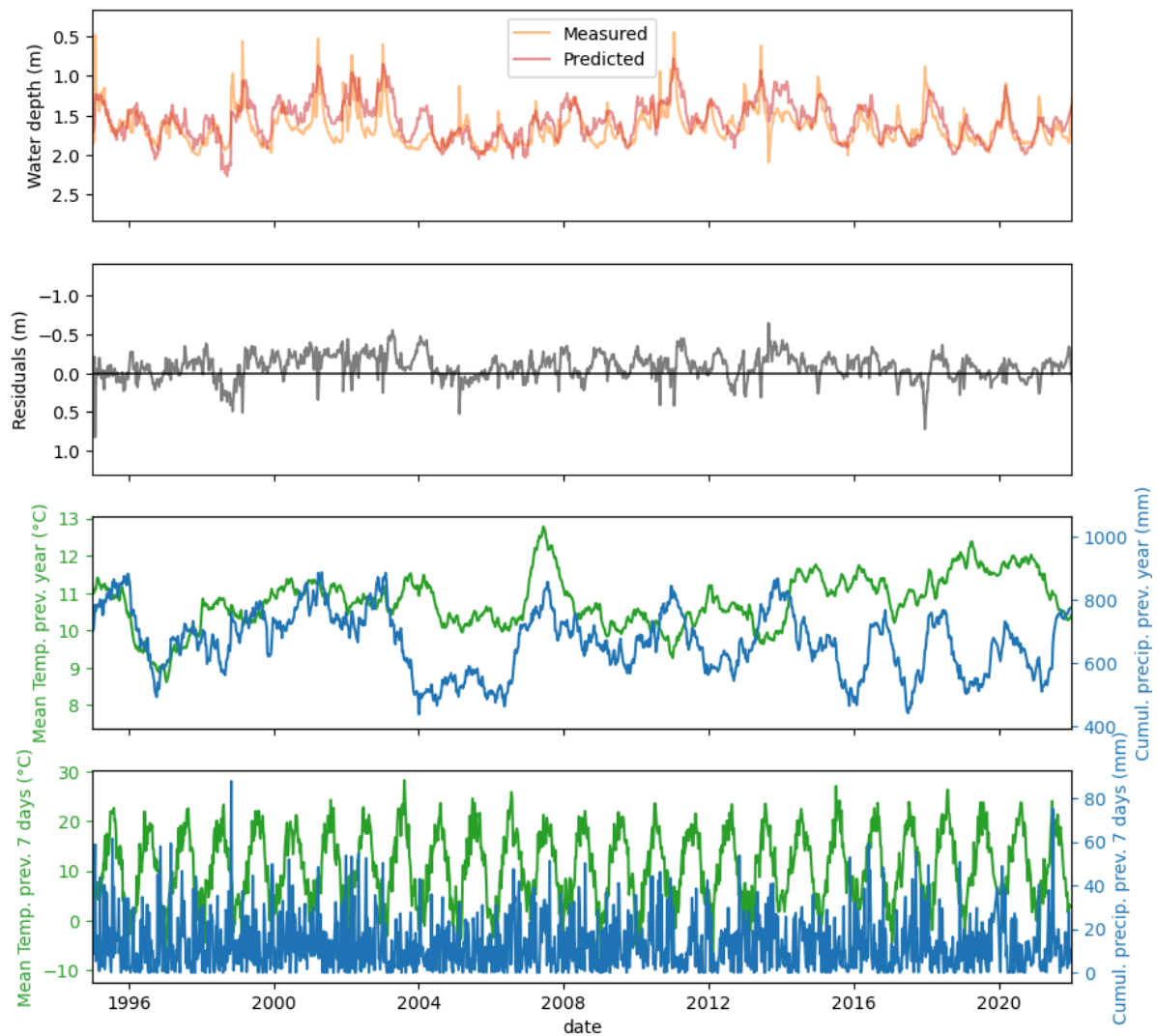- Saving the final trained model in an external file for later use.

Fig. 3. Model performance for a selected station from the test set for the period 1995-2022. The upper two panels show a good alignment of prediction and actual data (low residuals). The model captures the seasonal fluctuation well, and also the timing of extreme events (i.e. floods). Here, however, the magnitude of these short-lived excursions is usually underestimated (spikes in residuals to high values). The lower two panels show part of the data used to predict the groundwater levels, i.e., temperature and precipitation over different integration periods.
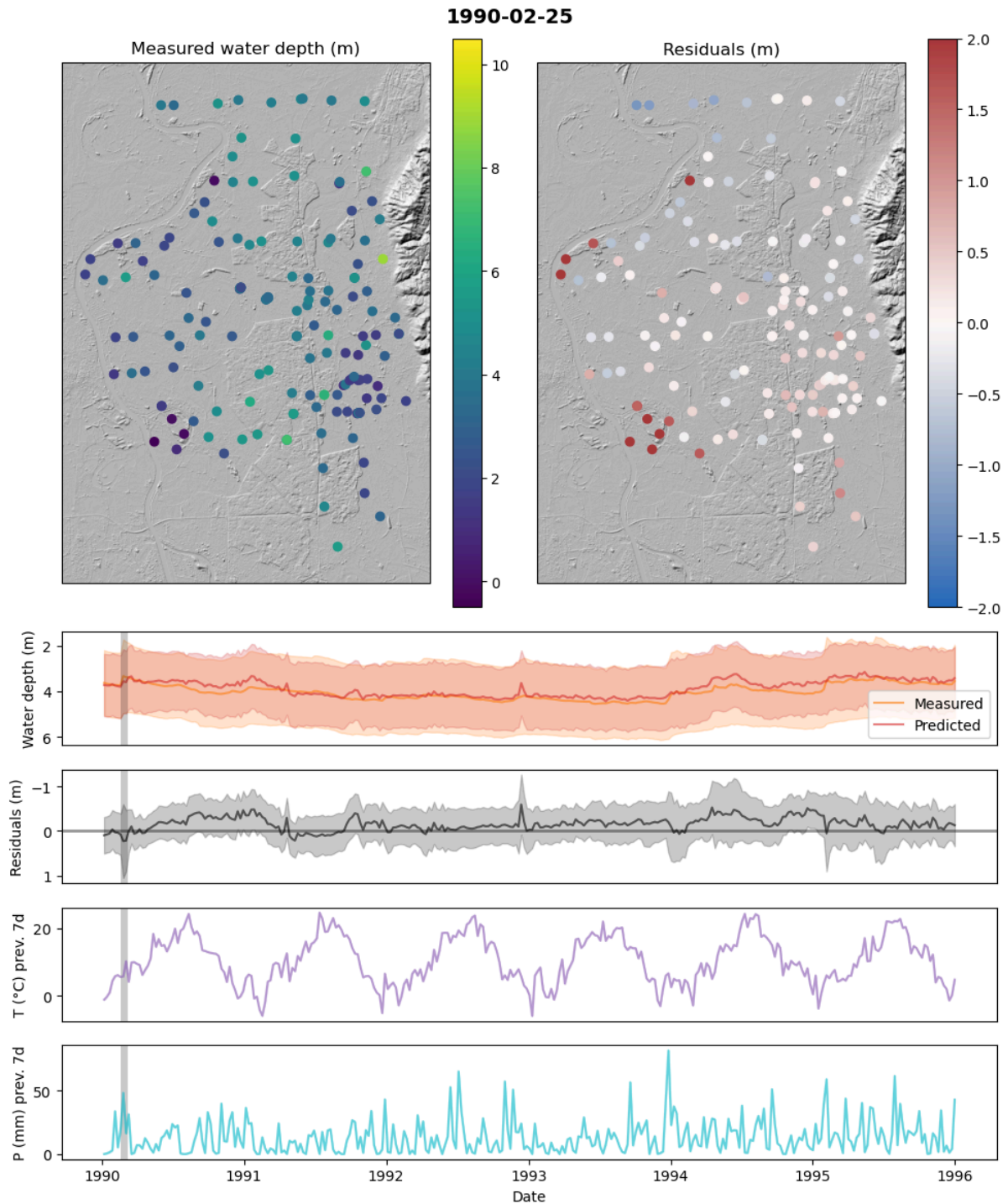
Fig. 4. Maps show measured and predicted (as residuals) groundwater levels for all stations for a day with a flood event along the Rhine river, which is not captured well by the affected stations (high residuals). Predicted vs. measured groundwater levels for all stations in the upper two time-series panels (solid line = all-station average; shaded region = all-station standard deviation) shows good overall agreement of the model with the actual data. The lower two time-series panels show selected meteorological parameters.

B3: Creation of future weather scenarios

The notebook model_weather.ipynb is used to create synthetic future weather scenarios, i.e. tables with daily or weekly entries for several years that contain weather data as it might look like in some future climate. As of now, the notebook contains two distinct and simple weather modeling approaches (addition and average models), but this could be improved substantially in the future. The synthetic weather data is finally stored in an external file for later use. The figure below illustrates the difference between "reference" weather and "addition" model weather in extreme climate conditions.
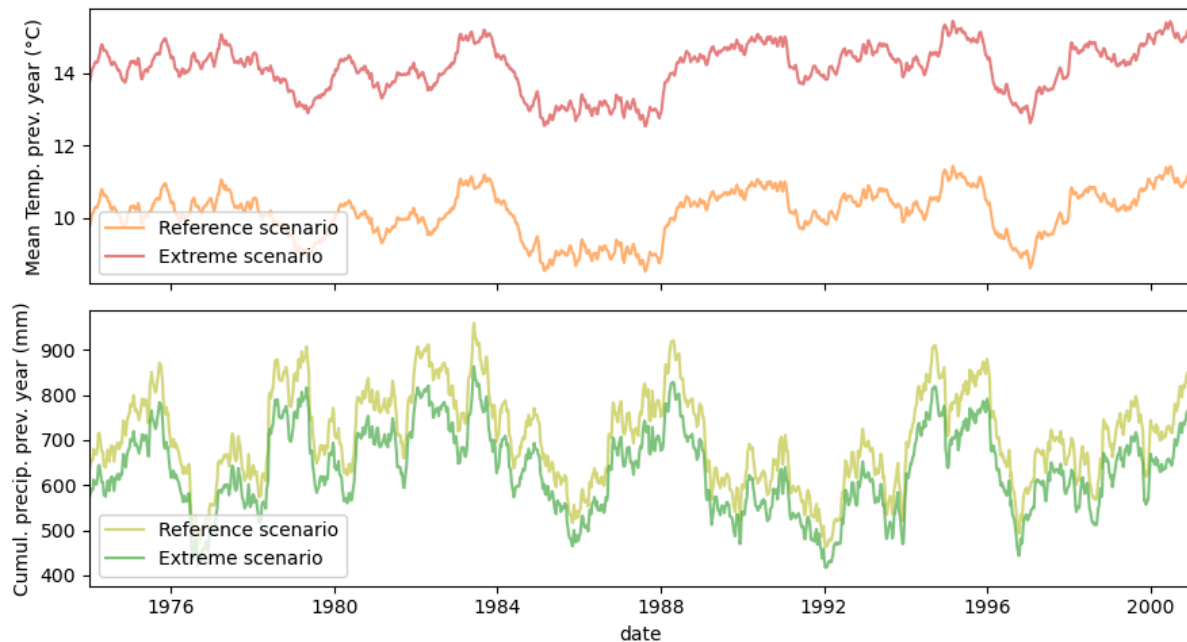

Fig. 5. A simple, synthetic weather model for a hypothetical extreme future climate, where in the period 2071-2100 (i.e. 100 years after the reference period) Temperatures are elevated by 4 °C and precipitation is reduced by 10%.

B4: Future groundwater prediction

The actual prediction of groundwater levels for each station and future weather is performed by the notebook future_model.ipynb. It loads the trained model, all needed groundwater station metadata and the weather scenarios. Then the model performs the groundwater prediction for all stations and its results are saved in an external file for further use.

B5: Analyze prediction results

The prediction results are analyzed in the notebook analyse_predictions.ipynb. The main parameter for comparing the effect of more extreme weather on groundwater levels is the groundwater anomaly, which is the difference in groundwater levels in reference vs. extreme weather for a certain time interval (see Fig. 6).
The maps and charts can be combined to animations using the functionality provided in notebook make_animations.ipynb.
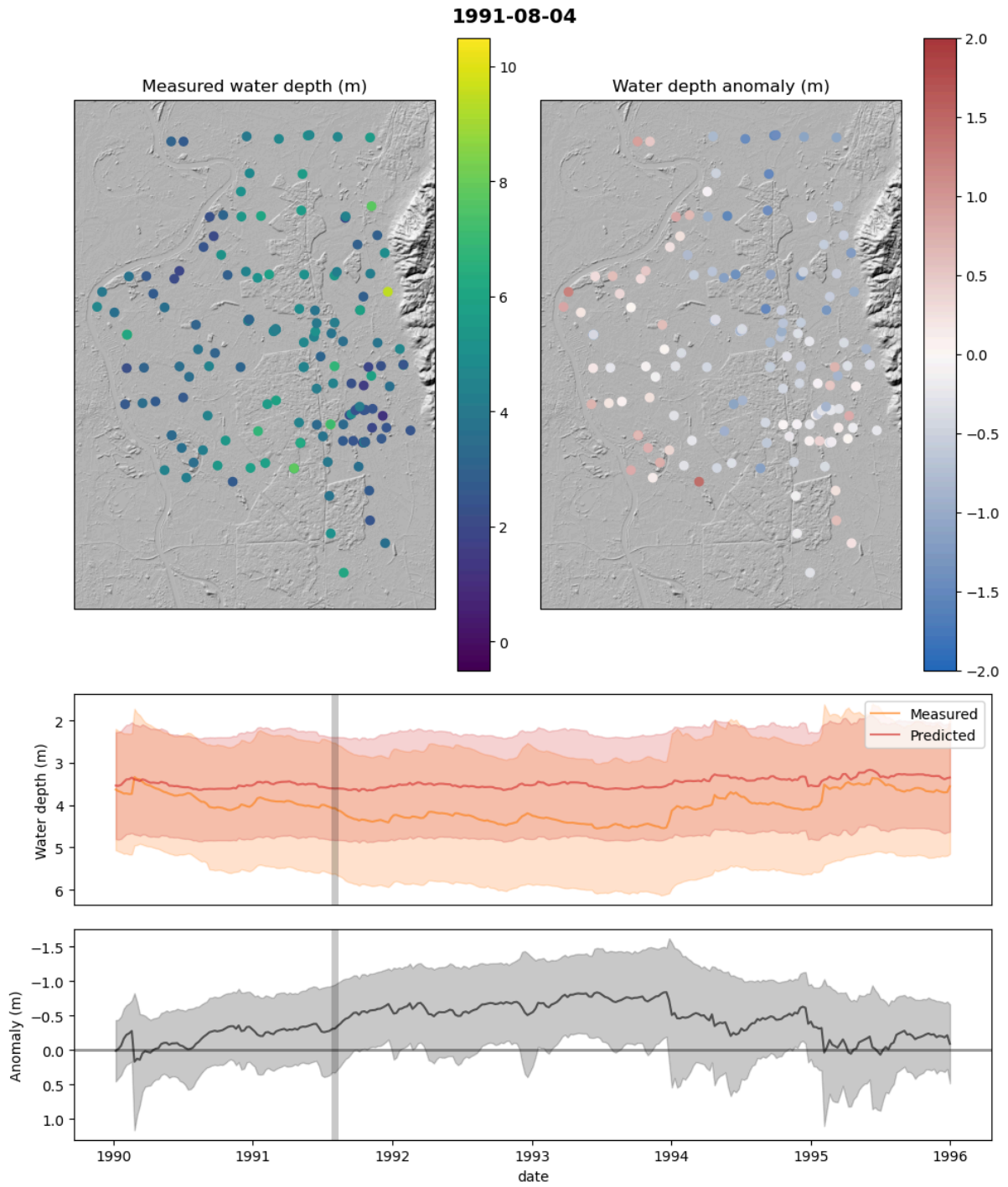
Fig. 6. Difference (anomaly) in groundwater levels for all stations on one selected day, when comparing the climate of the reference period with an extreme climate 100 years later.

C: Conclusion

Predicting past groundwater levels only using meteoric parameters works surprisingly well for the study area, even with a very crudely trained model. The results could potentially be improved even more when taking into account more meteoric parameters, and by using a more sophisticated training process (e.g. proper hyperparameter search), but this requires more computational resources than those available to the author.

Predicting groundwater levels for future climates using machine learning models might in principle be possible for the study area, at least to some degree. However, here the major drawback is the need for a realistic synthesis of future weather data, which appears to be a big challenge on its own.