An ensemble groundwater prediction (EGP) system to forecast groundwater levels in alluvial aquifers in Switzerland

Raoul A. Collenteur[1,2,3], Konrad Bogner[2], Massimiliano Zappa[2], Mario Schirmer[1,4], and Christian Moeck[1,5]

[1]Eawag, Department Water Resources and Drinking Water (W+T), Ueberlandstr. 133, Duebendorf, CH-8600, Switzerland
[2]Swiss Federal Research Institute WSL, Birmensdorf, Switzerland
[3]now at Collenteur Hydroconsult GmbH, Spiez, Switzerland (raoul.collenteur@hydroconsult.ch)
[4]Laval University, Quebec City, Quebec, Canada
[5]Swiss Groundwater Network (CH-GNet), Duebendorf, Switzerland

Correspondence: Raoul Collenteur (Raoul.Collenteur@eawag.ch)

Abstract. Groundwater is a key source of freshwater for drinking water supply and agricultural irrigation on a global scale. Groundwater in Switzerland (and beyond) is traditionally regarded as a reliable source of freshwater. Recent extreme drought events (i.e., in 2018, 2022, and 2026) have shown, however, that groundwater responds to these events and can cause problems for water supply and groundwater availability. With hydrological extremes becoming more frequent, there is a growing need for early warning systems and improved forecasting. This study develops and tests a scalable ensemble groundwater prediction (EGP) system with a 32-day lead time. The system combines extended-range precipitation and temperature forecasts from the European Centre for Medium-Range Weather Forecasts (ECMWF) with the lumped-parameter groundwater model Pastas. Forecasts were evaluated at six monitoring wells across Switzerland, representing diverse hydrogeological settings, and compared to naive persistence and climatology benchmarks. Results indicate that the EGP system produces skillful forecasts up to 32 days ahead, with Spearman correlations between 0.73 and 0.99 for one to four weeks ahead. The required model–data complexity to provide skillful forecasts varies. In aquifers with long response times and memory, forecasts driven by recent measurements of precipitation and temperature, as well as climatology, are sufficient. In fast-responding, short-memory systems, however, using meteorological forecasts of precipitation and temperature adds clear value and improves groundwater level forecasts compared with using only recent meteorological measurements and system memory. Forecast skill in mountainous regions (e.g., Davos) remains limited because local meteorology is difficult to predict. These findings highlight both the potential and the limitations of short-term groundwater forecasting.

1 Introduction

Groundwater is a key source of freshwater for drinking water supply and agricultural irrigation on a global scale, shaping ecosystems and landscapes (e.g., Rodell et al., 2018; Scanlon et al., 2023). Its importance is especially evident at regional scales, where it sustains local communities and economies. In Switzerland, about 80% of drinking water is drawn from groundwater (BAFU, 2022). Groundwater in Switzerland and beyond is traditionally regarded as a reliable source of freshwater, offering greater resilience to climate variability than surface water and other sources (Bordes et al., 2011). Recent extreme drought events (i.e., in 2018, 2022, and 2026) have shown, however, that groundwater does respond to these events and can cause problems in water supply and groundwater availability.

Switzerland has experienced an increase in extreme meteorological events, including droughts and heavy precipitation (e.g., Brunner et al., 2019; Scherrer et al., 2022; Tuel et al., 2022; Bauer and Scherrer, 2024). Scherrer et al. (2022) showed a clear tendency towards drier summers in Switzerland over the past four decades (1981–2020). Soil water content and climatic water balances reveal declining trends from spring to autumn, mainly driven by increasing evaporation and a slight decrease in precipitation. Future climate change is expected to increase variability in groundwater recharge and availability, including periods of low or high groundwater levels (Moeck et al., 2016; Epting et al., 2021) and decreased groundwater discharge in alpine headwater catchments (Halloran et al., 2023).

The rising frequency of hydrological extremes underscores the need for early warning systems and enhanced forecasting capabilities to better anticipate and manage their impacts on water resources. Hydrological forecasts can help water resource managers make more informed decisions, and decision makers can attempt to reduce the impact of meteorological extremes on water resources (e.g., White et al., 2017; Neumann et al., 2018). Possible measures include temporarily reducing pumping, enhancing managed aquifer recharge, or releasing water to mitigate flooding risks. Seasonal hydrological forecasts predicting groundwater levels up to a few weeks or months ahead may also be used to better inform groundwater users (e.g., drinking water suppliers and the agricultural sector) on possible limitations in groundwater extraction in the upcoming period (see, for example, Cantone et al., 2023).

Hydrological forecasts inherently come with a certain degree of uncertainty. To account for uncertainty, hydrological ensemble prediction is a common approach for generating forecasts of hydrological variables of interest (e.g., Duan et al., 2019). In this approach, a hydrological model is forced by ensembles of meteorological variables such as precipitation and temperature. The ensembles represent the uncertainty in the predicted meteorological forcings, and are themselves the output of a meteorological model. The forecasts used are typically issued for periods of weeks (short-term forecasts) to several months ahead (subseasonal forecasts), a time frame that is appropriate for operational water management. This sets this type of forecast apart from long-term climate change projections, which operate on much larger timescales and support the development of mitigation and adaptation strategies as well as long-term infrastructure investments, but do not provide the timely information needed for early warning or short-term response to individual extreme events.

Much of the hydrological ensemble prediction literature and applications focus on predicting streamflow (e.g., Duan et al., 2019; Wanders et al., 2019) and surface water levels in reservoirs (e.g., Viel et al., 2016), with fewer applications to forecasting

groundwater levels. One of the few examples of hydrological ensemble prediction of groundwater levels can be found in Mackay et al. (2015). This study found that skillful forecasts (i.e., predictions that perform better than reference baselines such as climatology or persistence forecasts) up to five months can be made for groundwater levels in the UK using the lumped-parameter groundwater model AquiMod. This system has also been operationalized (Prudhomme et al., 2017), showing that it is not merely an academic endeavor. Huang and Shih (2020) produced seasonal forecasts for Taiwan and concluded that predicted rainfall is crucial for groundwater level forecasts. In recent work, Robertson et al. (2024) introduced a new approach to simultaneously forecast streamflow and groundwater levels and successfully tested it in eastern Australia. A notable part of their approach is the use of an error model to reduce the forecast errors.

Despite extensive research on hydrological ensemble predictions in Switzerland (e.g., Bogner et al., 2018; Monhart et al., 2019; Padrón et al., 2025, and references therein), no system for predicting groundwater levels is yet available. Recent droughts have shown that groundwater in Switzerland is vulnerable to prolonged meteorological droughts, and that groundwater levels may respond within a few weeks to reduced recharge and lower river levels. This raised interest from governmental agencies in groundwater forecasting for droughts, for example, for use on the Swiss drought platform (https://www.trockenheit.admin.ch/en, last accessed 2026-08-27). Currently, this platform provides various hydrological forecasts (i.e., runoff, lake levels, and soil moisture) up to 32 days ahead, but the long-term goal is to extend the forecast period to 46 days. Even at shorter lead times, forecasts are expected to support decision-making to mitigate the impacts of groundwater droughts in faster-responding systems.

The goal of this study is to develop and test a scalable methodology to make ensemble groundwater predictions (EGPs) for alluvial aquifers in Switzerland. This is an initial step towards developing an operational groundwater forecasting system. Ensemble groundwater predictions are made using forecasts of precipitation and temperature from the extended range forecasts from the European Centre for Medium-Range Weather Forecasts (ECMWF) and the lumped-parameter model Pastas to simulate the groundwater levels (Collenteur et al., 2019). This study focuses particularly on forecast verification to investigate 1) if skillful forecasts of groundwater levels with lead times of up to 32 days can be made and 2) what combination of data and models is required for this purpose. The choice for a forecast up to 32 days ahead stems from the ECMWF data that were used, which provides meteorological forecasts up to 32 days ahead. To achieve this, the groundwater levels at six monitoring wells in Switzerland are modeled, and the forecasts from the EGP system are compared to three reference forecasts based on persistence and climatology. These six wells are selected based on their varying aquifer response times and the ability to model them with meteorological data only. This makes them relevant for addressing the question of what model-data combination is required to provide skillful forecasts.

2 Methods

2.1 General approach

The general workflow to simulate the groundwater levels and generate the ensemble groundwater predictions for each monitoring well is depicted in Figure 1. The flow path denoted with the blue arrows shows the steps to simulate the groundwater levels

using historic meteorological data. By iterating over this flow path (not shown) and changing the model parameters, the model is calibrated. The calibrated model is then used in forecasting mode (flow path with green arrows) to generate predictions of the groundwater levels using ensembles of the input data. Each of these steps is described in more detail below.

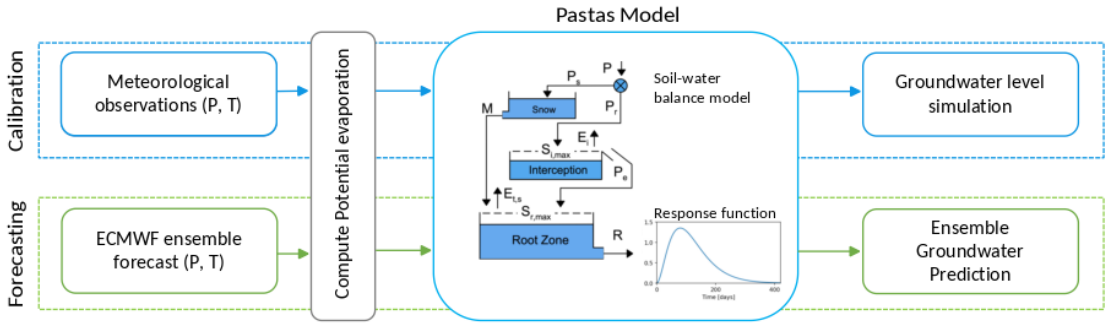

Figure 1. Schematization of the workflow applied in this study to simulate the groundwater levels (blue flow path) and generate the ensemble groundwater predictions (green flow path).

2.2 Groundwater level modeling

Lumped-parameter groundwater models from the Pastas software (Collenteur et al., 2019, version 1.14.0) are used in this study to simulate and forecast the groundwater levels. A schematic picture of the Pastas model is shown in the center of Figure 1. The groundwater levels are computed from the meteorological input data in a two-stage process. In the first stage, the groundwater recharge is computed from common daily meteorological input data (precipitation, potential evaporation, and temperature) using a soil-water balance approach (see Collenteur et al., 2021, for details). The soil-water balance model accounts for the effect of the limitation of actual evaporation due to soil moisture availability, and the temporary storage of infiltrated water in the root zone. This makes the groundwater recharge response (and ultimately the groundwater level response) to precipitation and potential evaporation nonlinear. The model also includes the effect of snowfall and snowmelt on groundwater recharge through a degree-day snow model (Collenteur et al., 2023). In the second stage, impulse response functions are used to describe the groundwater level response to impulses of recharge (von Asmuth et al., 2002). By convolving the groundwater recharge with the impulse response, a simulation of the groundwater levels is obtained. For each monitoring well, an individual model is developed to simulate and forecast the groundwater level at that location.

In Collenteur et al. (2023), the Pastas model was successfully applied to simulate historical groundwater level data and investigate stresses on the groundwater system at the same six locations used in this study. A distinct advantage of the Pastas model is that it runs fast and has short calibration times (in the order of seconds) even with limited computational resources. This is especially practical for ensemble predictions (Robertson et al., 2024), where many simulations are made to generate the forecast and include uncertainty from the Pastas model and the meteorological input data. For more detailed descriptions of the Pastas model, refer to Collenteur et al. (2021) and Collenteur et al. (2023).

2.3 Residual error modeling

The lumped-parameter model provides daily simulations of the groundwater levels for each monitoring well. When subtracting the simulation from the measurements, the residuals are obtained. Residuals from daily or weekly groundwater simulations typically show strong autocorrelation, which was also observed here. This violates a common assumption about the statistical properties of the model residuals, namely that the residuals are uncorrelated, normally distributed, and homoscedastic. Violating these assumptions may result in unreliable quantification of the parameter uncertainties, as, for example, shown in Collenteur et al. (2023). However, the existence of autocorrelation also indicates that there is still information in the residuals that may be used to further improve the groundwater level forecast. One strategy to deal with both issues, the one taken here, is to apply a noise model to model the residual errors. The commonly used autoregressive noise model of order one (AR1) is used in this study to model the residual errors (r_t) (von Asmuth and Bierkens, 2005):

$$r(t_i) = \upsilon(t_i) + r(t_{i-1})\mathrm{e}^{-\Delta t_i/\alpha} \tag{1}$$

where υ_t is assumed to be white noise, Δt_i is the time step between two residuals, and α is the decay parameter describing the relation between two consecutive residuals. This model adds one parameter (α) to be estimated. During model calibration, the noise model is used to transform the residuals into uncorrelated noise.

2.4 Model calibration

Eight model parameters must be inferred from the groundwater level data: seven from the lumped-parameter groundwater model and one from the aforementioned noise model. The parameters of the groundwater model and the noise model are inferred in a two-step process that was found to improve the estimation of the parameter (Collenteur et al., 2021). First, the groundwater model parameters are calibrated without the noise model. In the second step, these parameters are used as initial parameters for a second calibration, where the noise model is simultaneously calibrated. The sum of the weighted squared noise criterion, proposed by von Asmuth and Bierkens (2005), is used as the objective function for this second step. The objective function is minimized using a nonlinear least-squares approach. The Jacobian matrix computed by the optimization algorithm is used to estimate the covariance matrix and, ultimately, the standard errors of the parameters. These standard errors are subsequently incorporated to represent parameter uncertainty in the ensemble groundwater predictions.

Weekly groundwater level measurements from the ten years preceding the start of the forecast period are used for calibration. This period is preceded by a ten-year warmup period, for which meteorological data are available. Van der Spek and Bakker (2017) found that a 10-year calibration period yields good results in terms of groundwater level simulation and uncertainty quantification. However, this length was not optimized for the same goal as in this study: short-term prediction of groundwater levels. The model is recalibrated on the groundwater level measurements from the ten years before every forecast to include the latest available measurements. This ensures that the latest information on the state of the groundwater is considered when calibrating the model and generating the forecasts.

2.5 Generating the raw ensemble groundwater predictions

After the model parameters are estimated, the calibrated model is used in forecasting mode to generate the ensemble groundwater predictions in three steps. The model is forced with an ensemble of meteorological input data (with 51 ensemble members) and used to forecast the groundwater levels with a 32-day lead time. The number of ensemble members stems from the meteorological forecast data used in this study. Moreover, the forecast for each ensemble member is computed with 100 different parameter sets for every one of the six monitoring wells to include the effect of parameter uncertainty. The parameter sets are drawn from a multivariate normal distribution using the covariance matrix estimated during model calibration. This process generates 5100 raw groundwater level forecasts for each monitoring well (100 parameter sets multiplied by 51 ensemble members).

2.6 Post-processing the raw forecasts

The raw forecasts do not incorporate the information from the most recent groundwater level measurement before the start of the forecast. To improve accuracy, the raw forecasts are corrected using the fitted AR1 noise model and the last known residual, r_0, available before the start of the forecast at t_0. The correction g for a forecast at time t_h for lead time h is computed as:

$$g(t_h) = r_0 e^{-(t_h - t_{r_0})/\alpha}, \tag{2}$$

where t_{r_0} is the time of the last residual, and $h = t_h - t_0$ is the lead time in continuous time units (i.e., days). This correction is added to the forecasted groundwater levels. Its magnitude decreases with increasing lead time h, as the influence of the last residual r_0 diminishes over time.

2.7 Estimating the forecast variance and prediction interval

After generating the forecast, the forecast variance of the individual forecasts is estimated. Because the residuals are autocorrelated, the forecast variance increases with increasing lead times. Given the fitted AR1 noise model, the forecast variance (σ_h^2) for lead time h ahead can be computed as:

$$\sigma_h^2 = \sigma_v^2 \frac{\alpha}{2}(1 - e^{-2h/\alpha}) \tag{3}$$

where σ_v^2 is the variance of the noise. The forecast variance increases with increasing values of h, and converges to the variance of the residuals ($\sigma_r^2 = \sigma_v^2 \frac{\alpha}{2}$) as h goes to infinity. This approach assumes that the forecast errors behave similarly to the model residuals.

The steps described above produce an ensemble of 5100 forecasts and their individual forecast variances. The 95% prediction interval is computed under the assumption that the forecast ensemble follows a normal distribution at all lead times. The ensemble mean is calculated as the average of all individual members. The overall prediction variance is determined from

the ensemble members and the individual forecast variances using the law of total variance (see, for example, Blitzstein and Hwang, 2019):

$$\text{Var}_{\text{total}} = \text{Var}(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N) + \frac{1}{N}\sum_{i=1}^{N} \sigma_{h,i}^2, \tag{4}$$

where $\hat{y}_i$ is the groundwater forecast from the ensemble member i, $N = 5100$ is the number of ensemble members, and $\sigma_{h,i}^2$ is the individual forecast variance for the ensemble member i. Note that the forecast variance $\sigma_{h,i}^2$ varies due to uncertainty in the estimated parameter α. The standard deviation is then derived from this variance and multiplied by the critical value $z = 1.96$ to obtain the 95% prediction intervals.

2.8 Naive forecasts

To evaluate the performance of the ensemble groundwater prediction (EGP) system described above, the forecasts from the EGP system are compared to three naive forecasts. It is tested whether the EGP system outperforms other forecasting systems based on less data or simpler models, and thus provides any additional value over other or existing forecasting systems. The EGP forecasting system was compared to the following naive forecasts:

- **Persistence**: The persistence forecast is computed by extrapolating the last measured groundwater level over the forecast period, i.e., the last measured groundwater level is assumed to persist throughout the forecasting period. This approach is assumed particularly effective for aquifers with long response times, where groundwater levels evolve more gradually and persistence provides a reasonable baseline.

- **Autoregressive**: This forecast is created by fitting an autoregressive (AR) model on the groundwater level time series with weekly measurements. This model uses the groundwater level measurements from the past few weeks to predict the groundwater levels over the next 32 days. Autoregressive models have been extensively used to predict groundwater-level time series and consider trends from past weeks when generating forecasts. This approach is assumed effective for aquifers with relatively short response times, where recent trends in groundwater levels provide useful information for near-term predictions.

- **Climatology**: The climatology forecast is created by forcing the Pastas groundwater models with the mean precipitation, potential evaporation, and temperature for each day in the forecast period from the past 30 years (i.e., the average 30-year value for the 1st of January). This forecast is used to investigate the value of the meteorological forecast data. This approach is assumed particularly effective for aquifers with strong seasonal cycles, where long-term averages of meteorological conditions provide a reliable benchmark for expected groundwater dynamics.

The data-model complexity increases from the persistence forecast to the climatology forecast. The persistence approach requires only a single groundwater level measurement (the most recent), whereas the autoregressive method relies on the full

groundwater level time series to fit the model. The climatology forecast requires a lumped groundwater model and the average meteorological conditions over the forecast period. This contrasts with the EGP system, which requires a lumped groundwater model, the historic meteorological conditions before the forecast period, and the forecasted meteorological variables. In addition to the above forecast types, the climatology and EGP forecasts are run with and without post-processing the raw forecasts to investigate the effect of the correction on the forecast quality. With this setup, it can be tested whether the forecast skill is determined by the model, the historical data, or the forecasted meteorological input data.

2.9 Groundwater model evaluation

The model performance during the calibration period is evaluated using the mean absolute error (MAE) and the Nash-Sutcliffe Efficiency (NSE) metrics for goodness-of-fit. To evaluate the models, the initial models calibrated on the period 2002-2012 are used. Apart from computing these metrics on all the data in this period, the metrics are also computed on the different seasons to evaluate how model performances vary seasonally. This evaluation can be used to investigate the relationship between model performance during the calibration period and performance during the forecast period.

2.10 Forecast verification

An important aspect of developing an ensemble prediction system is assessing the quality of the generated forecast. Using past forecasts from 2012 to 2023 allows evaluation of the system. The forecast quality is evaluated using various verification metrics, as suggested in Duan et al. (2019), outlined below:

- **Correlation**: The Spearman correlation (R) between the mean of the forecast ensembles and the measured groundwater levels is computed and visualized to investigate how the mean forecasted heads are associated with the measured groundwater levels in the forecast period.
- **Accuracy**: The continuous ranked probability score (CRPS) is computed to evaluate the accuracy of the forecasts. The CRPS is a generalization of the mean absolute error to probabilistic forecasts. The lower the value of the CRPS, the closer the forecast is to the measured values. A value of zero would indicate a perfect forecast.
- **Skill**: The continuous ranked probability skill score (CRPSS) is computed to test if a model has skill (e.g., Hersbach, 2000; Gneiting et al., 2005). The skill scores for the CRPS metric are computed by comparing to the naive forecasts. The model is said to have skill if it outperforms these naive forecasts (CRPSS > 0) (Jolliffe and Stephenson, 2012). The scores are computed for each lead time (1-32 days) and over different months to distinguish in which period the forecast quality is better or worse.
- **Uncertainty (1)**: The spread-to-error ratio is computed to assess the uncertainty of the forecasts, similar to, for example, Monhart et al. (2019). The spread is the mean standard deviation of the ensemble forecasts, derived from the total forecast variance (Eq. 4). The error is measured as the mean of the root mean squared error (RMSE) of the ensemble means over all forecasts. The spread-to-error ratio thus compares the variability of a forecast ensemble (the spread) to the average

error of the forecast mean. For a forecast to be reliable, the spread-to-error ratio should be around one, indicating that the ensemble spread is approximately equal to the mean forecast error (Ho et al., 2013). A ratio below one indicates overconfidence (spread too narrow), while a ratio above one indicates underconfidence (spread too wide).

- **Uncertainty (2)**: The prediction interval coverage probability (PICP) is computed for all lead times and monitoring wells to further evaluate the quality of the prediction intervals. The PICP has a value between 0 and 1, where zero indicates that no measurements are within the prediction interval and 1 indicates that all measurements are in the prediction interval. The PICP target value is set to 0.95, as the 95% prediction intervals were estimated. Values above 0.95 indicate that the intervals are too wide, and values below indicate that the values are too narrow.

3 Data

3.1 Groundwater level data

Six monitoring wells from the Swiss national groundwater monitoring network (NAQUA) operated by the Federal Office for the Environment (FOEN) were selected for this study. These shallow wells are distributed across Switzerland, from north to south, and span different altitudes. This is reflected in the absolute groundwater level heights as well as in their diverse dynamics and response times to recharge, including snowmelt-driven processes (e.g., Davos). The locations of the selected monitoring wells and the measured groundwater levels are shown in Figure 2. In Collenteur et al. (2023) it was shown that the groundwater levels at these monitoring wells could be simulated with the lumped-parameter groundwater model Pastas with high accuracy, using only daily precipitation, potential evaporation, and mean air temperature as stresses. Based on this study and to the best knowledge of the authors, the groundwater levels at these wells are not influenced by other stresses such as pumping and surface water fluctuations. This aligns with the focus of the present study on ensemble groundwater predictions in natural systems, where groundwater level fluctuations are predominantly driven by meteorology.

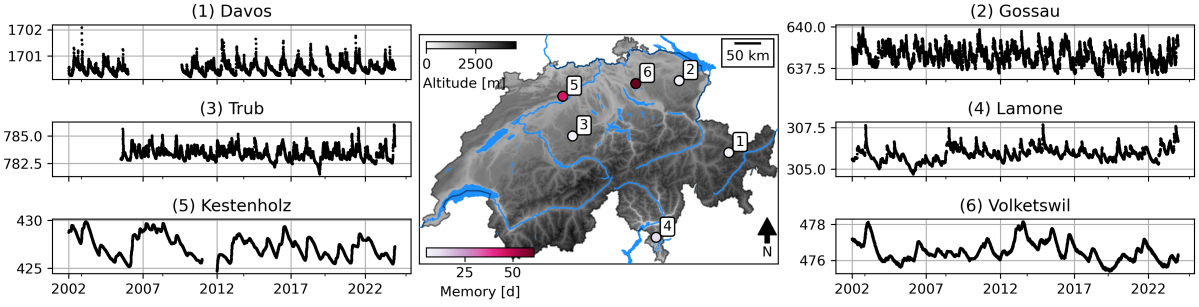

Figure 2. Locations and the hydraulic head time series of the monitoring wells in Switzerland. The color of each dot denotes the memory time (in days) of the groundwater system.

ID	index	Start	DTW	Alt.	Prec.	Evap.	Temp.	Snow	Memory
1	Davos	1990	2.0	1703	1057	379	2	94	5
2	Gossau	1990	7.0	645	1331	562	9	26	7
3	Trub	2005	8.0	791	1494	505	7	45	7
4	Lamone	1990	4.7	311	1709	647	12	3	13
5	Kestenholz	1990	19.1	447	1104	575	10	19	45
6	Volketswil	1990	38.6	515	1185	586	10	17	61

Table 1. Overview of some characteristics of the groundwater monitoring wells. DTW = average depth to water table [m], Alt = altitude in meters above sea level, Prec = average annual precipitation [mm/a], Evap = average annual potential evaporation [mm/a], Temp = average air temperature at 2 meters above the land surface [$^{\circ}$C], Snow = the average number of snow days, and the memory is the estimated memory of the groundwater system in days. A snow day is defined as a day with precipitation above zero and temperature below zero$^{\circ}$C. All the meteorological characteristics are computed over the period 1993-2023.

Table 1 provides an overview of the characteristics of each monitoring well. All wells are located in alluvial, unconfined aquifers. The depth to the water table ranges between 2 and 38.6 meters. The groundwater systems have a wide range of memory times, meaning they respond to meteorological inputs on different time scales. In this study, the memory time is defined as the first time lag (in days) where the autocorrelation in the groundwater level data drops below or equals 0.9. The wide range in memory times, from 5 to 61 days, allows investigating the relationship between this characteristic of groundwater systems and the forecast quality and skill.

The groundwater level data is available since 1990 as part of an active groundwater monitoring network. Only for the well in Trub the measurements start in 2005. Another important characteristic of these monitoring wells is that the loggers are connected via telemetry, and the water level data is continuously transmitted to a database. This allows the forecaster to improve forecasts using the latest groundwater level measurements in a potential operational setting. As mentioned earlier, assimilating the latest groundwater level measurements substantially improved forecast quality in the preliminary phase of this study; therefore, monitoring wells connected to telemetry were selected for this study.

3.2 Meteorological data

Historical meteorological data for the period 1990-2023 is obtained from gridded precipitation and air temperature data sets (RhiresD and TabsD, respectively) provided by MeteoSwiss (MeteoSwiss, 2022). For each monitoring well, the daily values from the grid cell in which the well is located are used. The grid spacing is approximately one kilometer, but it is noted by MeteoSwiss that the effective resolution is in the order of 15-20 km. Potential evaporation is computed from the mean daily air temperature using the Hamon equation as implemented in the Python package PyEt (Vremec et al., 2024). Table 1 also provides summary statistics of the meteorological conditions at each monitoring well. Annual average precipitation ranges

from 1057 to 1709 mm per year in Davos and Lamone, respectively. The average annual potential evaporation ranges between 379 mm per year in Davos and 647 mm per year in Lamone. Particularly the wells in Trub and Davos experience many snow days every year, with substantially lower air temperatures.

3.3 Forecast data

Extended-range forecasts of precipitation and temperature between 2012 and 2023 with a lead time of up to 32 days from the extended-range forecasts from the ECMWF Integrated Forecasting System (IFS) are used in this study. Forecast data issued once per week are available for the period 2012-2018, and forecasts issued twice per week are available for the period 2018-2023. Each forecast contains ensembles of 51 members for precipitation and temperature. The potential evaporation is computed using the same methods as for the historic meteorological data, as outlined above. These (raw) meteorological forecasts are available with a spatial resolution of $\sim$ 30 km and have been downscaled to 2 km using a thin plate spline regression, a non-parametric surface estimation method (Wahba, 1990). For more details on the meteorological forecast data, the reader is referred to Monhart et al. (2018). In total, 875 forecasts are available for the period of interest and used to evaluate the groundwater level forecasting system. The data is available in a gridded format. As with the historic meteorological data, values from the grid cell where the monitoring well is located are extracted.

It was explicitly chosen not to bias-correct the meteorological forecast data. The primary reason is that MeteoSwiss would perform this task in the system's foreseen operational setting. Additionally, bias-corrected forecasts are only available from 2018 onward, lowering the number of (re)forecasts that would be available for the forecast verification. Bias correction of the meteorological forecast data is expected to have the largest positive effect in mountainous regions dominated by snowmelt processes (Monhart et al., 2019). Except for Davos, all groundwater stations are located in the lower elevations, where bias correction showed only minor improvements compared to post-processing of hydrological forecasts (Bogner et al., 2022).

4 Results

4.1 Groundwater model performance

Table 2 shows the model performance metrics for the initial calibration period 2002-2012. The metrics are computed on daily groundwater level measurements, for all six monitoring wells and four different seasons. Visualizations of the measured and simulated groundwater levels, as well as the autocorrelation plots for the six wells, are available in Appendix B. The autocorrelation plots show that for five out of six models, no significant autocorrelation exists in the noise (υ in Eq. 1). Only for the model for Kestenholz is there significant autocorrelation in the noise, indicating that the estimated parameter uncertainties may be incorrect. The forecast uncertainty due to parameter uncertainty should thus be treated with caution for this well.

The average MAE ranges between 0.07 and 0.21 meters for Davos and Gossau, respectively. The average NSE values range between 0.71 for Gossau and 0.97 for Kestenholz. Depending on the monitoring well, model performances may differ substantially between the different seasons. Particularly for the monitoring well in Davos, the model performs better in spring

(MAM) and autumn (SON) compared to winter (DJF) and summer (JJA). Especially the winter is strongly influenced by snow accumulation and melting, and is crucial for the simulation of the groundwater levels. For the other wells, the groundwater levels measured in the spring (MAM) and summer months (JJA) are more challenging to simulate compared to the winter months, potentially due to the influence of vegetation growth and processes important for evapotranspiration rates. Models for wells with long memory times (i.e., Kestenholz and Volketswil) tend to perform better than models for wells with shorter memory times (i.e., Davos and Gossau). Overall, the models can accurately simulate the groundwater levels, a prerequisite to using the models for the prediction of the groundwater levels.

		AVG	DJF	MAM	JJA	SON
MAE	Davos	0.07	0.05	0.10	0.09	0.06
	Gossau	0.17	0.15	0.22	0.18	0.15
	Trub	0.15	0.16	0.22	0.14	0.11
	Lamone	0.20	0.19	0.21	0.19	0.19
	Kestenholz	0.22	0.21	0.19	0.21	0.26
	Volketswil	0.11	0.14	0.12	0.07	0.09
NSE	Davos	0.76	0.27	0.74	0.51	0.67
	Gossau	0.80	0.81	0.72	0.77	0.85
	Trub	0.82	0.81	0.68	0.90	0.85
	Lamone	0.64	0.64	0.61	0.65	0.65
	Kestenholz	0.96	0.97	0.97	0.95	0.95
	Volketswil	0.90	0.92	0.91	0.91	0.84

Table 2. Goodness-of-fit metrics for the different seasons and the entire year (AVG) of the models calibrated in the period 2002-2012. MAE is the mean absolute error in meters, and the NSE is the Nash-Sutcliffe efficiency [-]. AVG = average over the year; DJF = December, January, February; MAM = March, April, May; JJA = June, July, August; SON = September, October, November.

4.2 Examples of the ensemble groundwater predictions

Figure 3 shows four examples of the resulting ensemble groundwater predictions for two different start times (one in spring (left) and one in fall/winter (right)) and two monitoring wells: Gossau (above) and Kestenholz (below). The black dots denote the measured groundwater levels, and the blue line denotes the mean groundwater level forecast. The actual forecasts start at the vertical black dashed line. The gray lines indicate the mean of the 100 forecasts with different parameter sets, for each of the 51 ensemble members stemming from the meteorological forecasts. The shaded gray area denotes the 95% prediction interval over the entire forecast ensemble. This indicates that most prediction uncertainty stems from the meteorological forecast data rather than model parameter uncertainty. This trend was consistently observed across all forecasts.

The groundwater system in Gossau has a much shorter memory time than Kestenholz (7 vs. 45 days), as shown by more rapid fluctuations. The prediction intervals for Gossau show a wider range, indicating that the forecast is more uncertain compared to Kestenholz. The forecast range for Kestenholz is considerably narrower, and almost negligible during fall/winter. This is due to the long memory time of the system and the corresponding buffer capacity. For the well in Kestenholz, forecasts depend less on meteorological forecasts and more on system memory, whereas the effect of meteorological forecast data is much more visible in the forecasts for Gossau. Forecast quality depends more on input data in spring and summer, when groundwater levels are rising. In fall/winter, when groundwater levels are dropping, the forecast quality depends more on the memory of the groundwater system and less on the input to the system. The mean forecasted groundwater levels align well with the observations, a consistency that is also reflected in numerous individual forecasts. The figures for all the individual forecasts and wells can be found in the data repository (see the data and code availability section).

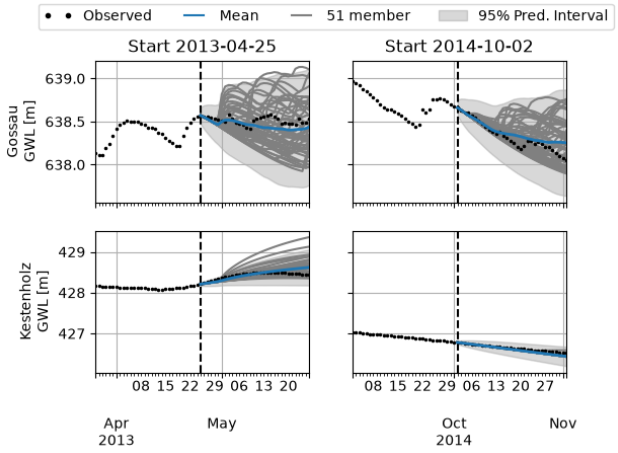

Figure 3. Two examples of forecasts with two starting dates for the monitoring wells in Gossau and Kestenholz. The black dots denote the measured groundwater level, the blue line the mean forecast, and the gray lines represent the mean of 100 forecasts for each of the 51 ensemble members. The dark shaded area shows the 95% prediction interval of the forecast.

4.3 Forecast quality

4.3.1 Predicted versus measured groundwater levels

Figure 4 shows the ensemble mean prediction (y-axes, the blue lines in Figure 3) against the measured groundwater levels (x-axes) at lead times of one to four weeks (each row is one lead time), for each of the monitoring wells (each column). The monitoring wells are ordered from short (left) to long (right) memory times. The orange line is the 1:1 line; if all the dots are on this line, it indicates a perfect forecast. The ensemble mean and measured groundwater levels generally show a high correlation, indicated by Spearman correlation coefficients, all exceeding 0.73. The forecasts for Davos, Trub, and Lamone show a small consistent bias, where the low groundwater levels are overestimated and the high groundwater levels are underestimated. For

the remaining wells, this is not the case. For the well in Davos, the high groundwater levels show a relatively large spread despite a strong correlation coefficient.

For the well in Trub, forecasts show high accuracy at short lead times (R = 0.95 for one week), but performance steadily declines as the forecast horizon increases, reaching R = 0.78 at four weeks. At Trub, the forecast-observation relationship departs from the 1:1 line at the 4-week horizon, with high groundwater levels underestimated and low levels overestimated. This pattern indicates a damping of the extremes in the forecasted groundwater levels. For the well in Volketswil, forecast performance remains high across all lead times, with correlation values close to one and points tightly aligned along the 1:1 line. This stability reflects the aquifer's long response time, where groundwater levels change gradually and are less sensitive to short-term errors in the meteorological input. As a result, even at longer forecast horizons, the predictions capture both the magnitude and dynamics of the measured groundwater levels with very little loss of skill.

Overall, the forecast quality (measured as the Spearman correlation R) decreases with increasing lead times for all wells, as expected, but remains high even at a lead time of four weeks. The data show that groundwater systems with larger memory times (i.e., Volketswil and Kestenholz) generally have a higher forecast quality than faster-responding systems (i.e., Davos and Gossau). This indicates that groundwater levels in systems with longer memory times can be predicted more accurately. In contrast, in more dynamic systems, even small errors in the input data can propagate quickly and become visible in the forecasts, reducing predictive performance.

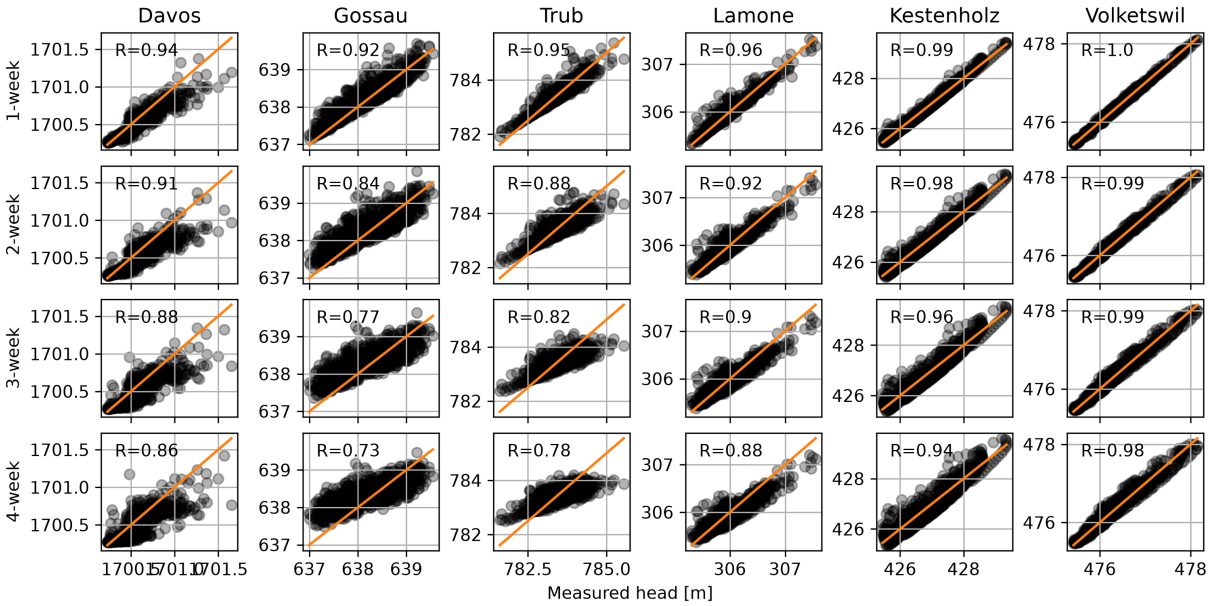

Figure 4. Correlation plots for each monitoring well (each row) of the measured heads (x-axes) against the forecasted heads (y-axes) at lead times of one, two, three, and four weeks. Each dot represents the mean value from the ensemble prediction. The orange line indicates the 1:1 line.

4.3.2 Continuous ranked probability score

Figure 5 shows the average continuous ranked probability scores (CRPS) over all forecasts for the six monitoring wells (each subplot) and each forecasting method. The dashed lines denote the climatology and EGP forecasts when no post-processing is applied for comparison. Note that the scales of the y-axes differ between the different subplots. For all wells and forecasting methods, the CRPS increases with increasing lead time. indicating a loss in the forecast quality with increasing lead times. Overall, the persistence forecast performs worst, except for small lead times of 1-2 days. This is expected, as groundwater levels typically change gradually due to the slow response of aquifer systems, making the groundwater level tomorrow often very similar to today's.

The climatology and EGP-based forecasts generally show the lowest CRPS values and therefore the best performance. The EGP system outperforms the naive forecasts for the wells in Gossau, Trub, Lamone, and Kestenholz (at larger lead times). For the wells in Davos and Volketswil, the EGP and climatology forecasts show similar performance. For these wells, and Kestenholz for shorter lead times, it can be concluded that the meteorological forecast data does not add value, whereas for the other wells the data improves the forecast. The use of a lumped groundwater model and historic meteorological data is preferred over an autoregressive model, because the latter is outperformed by the former for most wells and lead times.

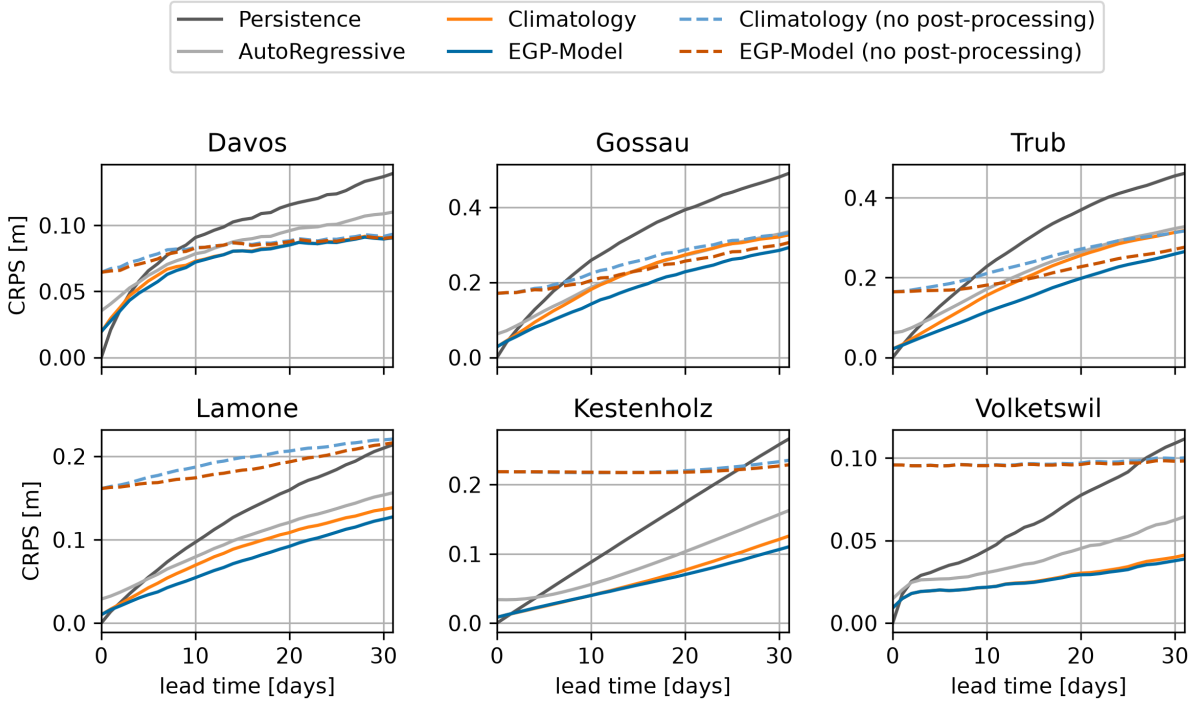

Figure 5. Continuous ranked probability scores (CRPS) for all four forecasting methods and all six wells. A larger CRPS indicates worse forecasts.

The comparison of the CRPS for the climatology and EGP forecasts with and without post-processing shows the value of the adding the correction. Using the last available groundwater level measurement before the start of the forecast substantially lowers the CRPS values for both forecasts. This is particularly the case for shorter lead times, where even the persistence forecast is better. Only for larger lead times do the forecasts without the post-processing outperform the persistence forecasts. Since post-processing clearly improves the forecast quality, only the post-processed forecasts are used in the remainder of this manuscript.

4.3.3 Continuous ranked probability skill score

Figure 6 shows the skill of the EGP system, measured as the CRPSS, compared to the persistence forecasts (left column), the autoregressive model (middle column), and the climatology forecasts (right column). The monitoring wells are ordered from short (top) to long (bottom) memory times. Each row shows the mean CRPSS for a given month across all forecasts. Values above zero (blue colors) indicate that the EGP system performs better, whereas values below zero (red colors) indicate that the EGP system performs worse. Values around zero (yellow colors) indicate no difference between the systems. The following general trends are observed in Figure 6:

- The positive CRPSS values in the left column indicate that the ensemble predictions (EGP) generally add value compared to persistence forecasts. For small lead times (i.e., forecasting a few days ahead), the added value is minimal, or the persistence forecast even performs better. This is particularly evident in winter at the Davos well, where the model struggles to capture the rapid and snowmelt-driven groundwater dynamics, resulting in lower skill.
- The positive CRPSS values in the middle column indicate that the EGP model outperforms the forecasts generated with an autoregressive model. This indicates that the use of the lumped-parameter model, in combination with meteorological input data, improves the forecast compared to a model solely based on the head data. The only real outlier is the well with the shortest memory time (Davos), where the autoregressive model performs similarly or even slightly better.
- The data shown in the right-hand column show that the meteorological forecast data do not necessarily elevate the forecast skill for all lead times and wells. Only for larger lead times and groundwater systems with shorter memories (except Davos) do the meteorological forecasts add value. For monitoring wells with shorter memory times, the EGP system outperforms the climatology forecasts for lead times larger than a few days. This indicates that using meteorological forecasts to force the models improves the groundwater level forecast for these wells. For the wells in Lamone and Volketswil, with the longest memory time, no or little value is added by the meteorological forecast data.

Seasonal patterns in forecast skill are clearly visible, indicating that prediction quality depends on the time of year and cannot be assumed constant. For Davos, the EGP system performs similarly to, or in some cases worse than, the climatology forecast, especially during winter, likely due to limitations in the meteorological forecast data (discussed later). In contrast, the data show more consistent added value of the EGP system for the well in Gossau across seasons. These examples highlight

that the skill of the ensemble groundwater prediction system is strongly site- and season-dependent, reflecting differences in aquifer response dynamics as well as the reliability of meteorological inputs.

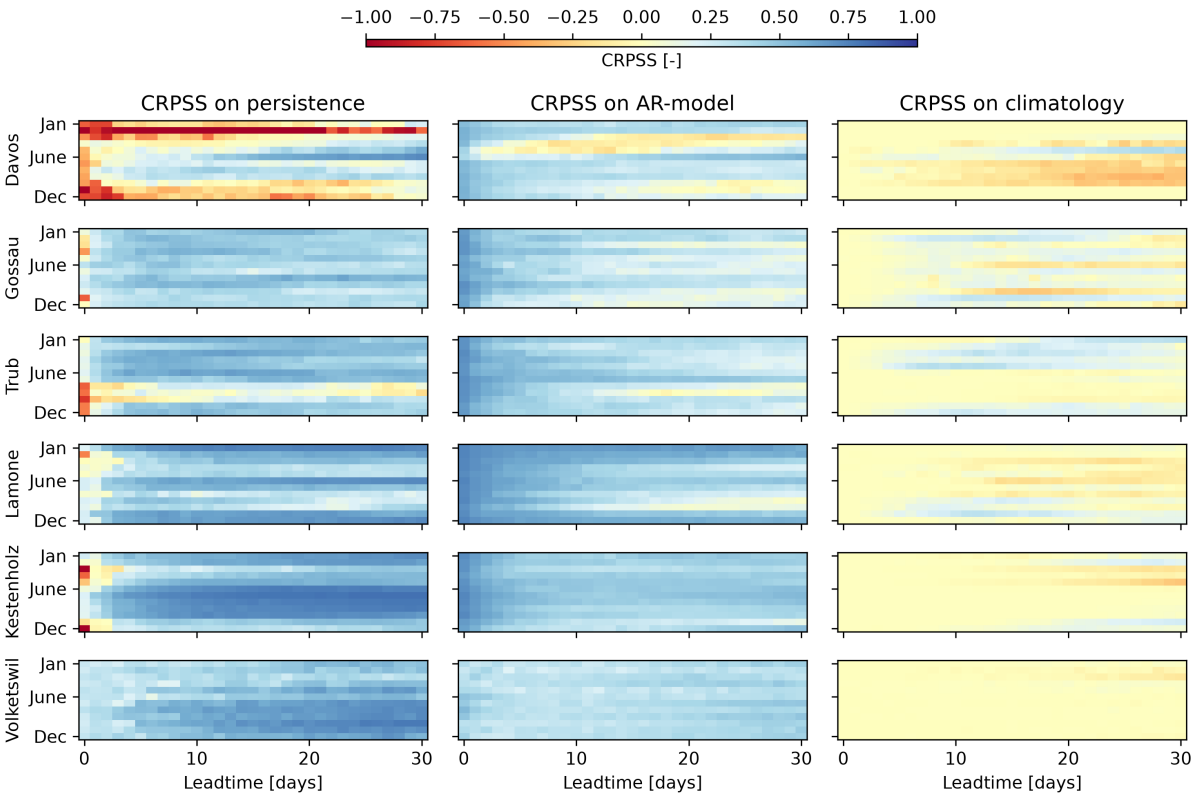

Figure 6. The forecast skill, measured as the CRPSS, for all monitoring wells (each row is one well). The EGP forecasts are compared to the persistence forecast in the left columns, to the autoregressive forecast in the middle column, and to the climatology forecast in the right column. CRPSS values above zero (blue colors) indicate that the EGP forecasts perform better, whereas values below zero (red colors) indicate that the EGP forecasts perform worse.

4.4 Evaluation of prediction intervals

An important aspect of a (probabilistic) forecasting system is the estimated uncertainty of the forecasted groundwater levels. Using past forecasts allows evaluation of the prediction intervals by comparing them with groundwater level measurements. Figure 7 shows the spread-to-error ratio (upper plot) and the PICP (lower plot) for all six monitoring wells. For the metrics shown in both plots, the target value is denoted by the dashed black horizontal line. Values above these lines indicate under-confidence (i.e., the prediction intervals are too wide), while values below the line denote overconfidence (intervals are too narrow).

Except for the smaller lead times (< ∼ 5 days) and the monitoring well Volketswil, the forecasts are generally slightly overconfident. The prediction intervals (i.e., the dark shaded areas in Figure 3) should thus ideally be slightly wider. Only for the well in Volketswil the forecasts are underconfident for larger lead times (∼ 1 week), as is most clearly visible from the PICP values for this well in the lower plot. A more in-depth analysis of the changes in PICP values through the seasons (shown in Figure C in the Appendices) reveals that the quality of the prediction intervals for Davos, Trub and Lamone shows strong seasonal variations. These could be linked to heteroscedastic residuals (i.e., the error is smaller at low groundwater levels and vice versa), which are not accounted for in the noise model. These may be caused by seasonal processes such as snowmelt-driven recharge (i.e., in Davos) and large summer precipitation events (i.e., in Lamone), which introduce variability and more abrupt groundwater responses that are not fully captured by the model and forecast data.

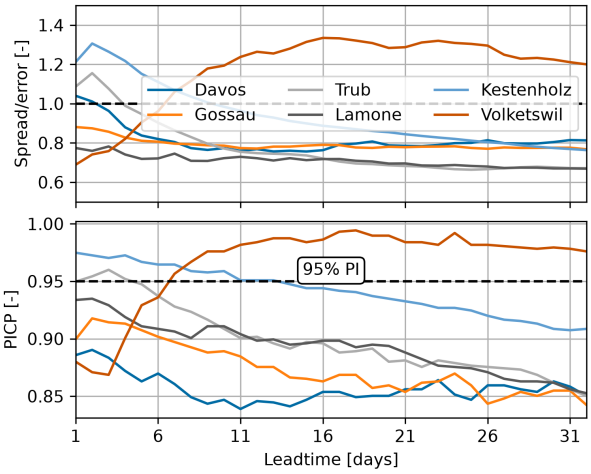

Figure 7. The spread to error ratio (upper plot) and the prediction interval coverage probability (PICP, lower plot) for all six monitoring wells. The dashed black lines denote the target value, when the uncertainty would be perfectly estimated according to each metric.

5 Discussion

5.1 Sources of forecast skill and memory time

The results shown in Figures 5 and 6 indicate that for small lead times (typically up to a few days), the meteorological forecasts added little to no value over forecasts with climatological input data. Moreover, for groundwater systems with longer memories and slower responses, i.e., Kestenholz and Volketswil, the meteorological forecasts add little to no value compared to the climatology forecasts for lead times of up to 32 days. Both forecasts made with the lumped-parameter model and meteorological measurements preceding the forecast period did perform substantially better than forecasts based on persistence and autoregressive models. This finding is in agreement with that of Mackay et al. (2015) for a similar study in the UK. Thus, for short lead times and long-memory systems, using meteorological forecast data does not add much to forecast quality

compared with a forecast based on climatology. The use of a model and the meteorological conditions before the forecasting period does (substantially) improve the forecast compared to methods that only use the preceding groundwater levels as data.

The results shown in Figures 5 and 6 also showed that for larger lead times (upwards from a few days) and for shorter memory systems (i.e., Gossau, Trub, and Lamone), the forecasts based on meteorology start outperforming the climatological forecasts. In these cases, the use of a model and meteorological forecast data clearly improves the forecast. The predictive skill can thus vary substantially across aquifers with different response characteristics. This finding is significant because it means that depending on the lead time and characteristics of the groundwater system (i.e., memory time), one can use different levels of model/data complexities to generate accurate forecasts. For systems with long memory times (i.e., Volketswil) and for lead times up to a few weeks, there appears to be no need to add the complexity of meteorological forecast data. For shorter memory systems and larger lead times, however, the meteorological forecast data does add value and improves the forecasts, particularly during the recharge period in winter (see Fig. 6).

An important source of the forecast skill comes from the last groundwater level measurement before the start of the forecast, as shown by the results in Figure 5. Incorporating information from this measurement is vital to produce high-skill forecasts at shorter lead times. This finding is in agreement with similar findings from Robertson et al. (2024), who used a comparable approach to post-process the raw groundwater level forecasts. Post-processing is particularly important when the focus is on the absolute value of forecasted groundwater levels, rather than just their dynamics. Post-processing mainly affects absolute groundwater level values, not their trend (e.g., decline or rise). Finally, this finding has implications for real-world systems, showing that choosing wells with telemetry and thus up-to-date measurements can help in providing better forecasts.

Given the lag in groundwater responses to meteorological variables, forecasts with larger lead times than those tested here might further benefit from meteorological forecast data. Future studies can test if groundwater level forecasts with larger lead times (up to a few months ahead) are feasible. Robertson et al. (2024) showed the potential for forecasting groundwater levels several months ahead in Australia, and it would be interesting to investigate whether this can also be extended to Switzerland and other regions with different environmental conditions.

5.2 Biases in meteorological forecasts

The low performance in the ensemble groundwater forecasts for some wells can in part can be explained by (consistent) biases in the meteorological forecasts. Figure 8 shows bias factors for the temperature and precipitation, respectively. The bias factor for the temperature was computed by dividing the average forecasted temperature by the average measured temperature for each forecast period. The bias factor for precipitation was computed by dividing the forecasted cumulative sum over the forecast period by the measured cumulative sum. Generally, the meteorological forecasts show relatively low or moderate biases for most of the wells. For the well in Davos, however, clear structural biases appear, probably due to its location in a mountainous area. In particular, the winter temperature bias substantially affects the representation of snow accumulation and snowmelt in the degree-day model. Even small systematic temperature deviations can shift the timing and magnitude of snowmelt, leading to errors in recharge estimates and, consequently, in simulated groundwater dynamics during winter and spring. This mechanism likely contributes to the lower CRPSS values for Davos in these seasons.

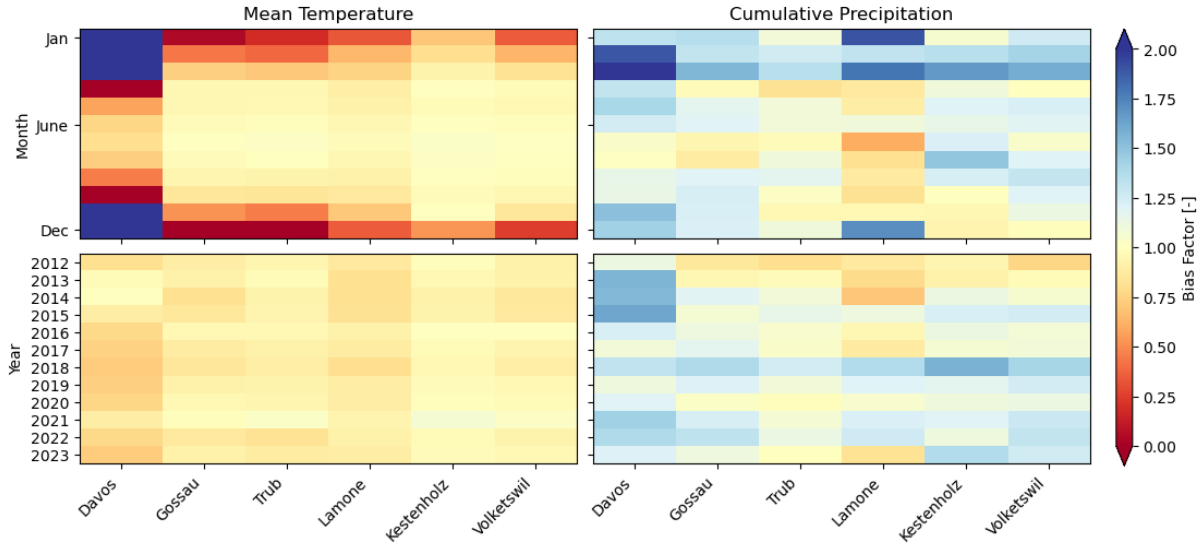

Figure 8. Bias factors for the temperature (left column) and precipitation (right column) forecasts, split out per month (top row) and year (bottom row). A bias factor below/above one indicates that the temperature or precipitation is underestimated/overestimated.

One possible solution is to bias-correct the meteorological forecasts. This was not implemented in this study, however, because bias-correction of the meteorological input data would be done as a separate step in the foreseen operational setting (see also Section 3.3). Moreover, bias correction in an operational setting may not be so straightforward and is outside the scope of this study. First, a long forecast period is required to determine the bias-correction parameters. Second, it needs to be assumed that the (seasonal) errors are structural and do not change between the years. The data shown in Figure 8 indicate that any correction factors, particularly for precipitation, are not constant through time. Nonetheless, it is expected that bias-correction of the meteorological forecast data will further improve the forecast quality over the results reported in this study.

5.3 Recommendations for future research

In this study, groundwater levels were simulated that resulted from changes in precipitation, air temperature, and potential evaporation. Many groundwater systems, including in Switzerland, are also impacted by other natural or anthropogenic stresses, such as surface water interaction and discharge changes, irrigation, and groundwater pumping. An advantage of the lumped-parameter model used in this study is that these stresses can be easily implemented in the model (von Asmuth et al., 2008). However, the difficulty in including these stresses may be in obtaining the input time series. This challenge applies to any model type and approach, as the accuracy of the forecast ultimately depends on the availability and quality of input data. Anthropogenic stresses, such as pumping, are often incompletely recorded and may need to be reconstructed from historical data. Alternatively, pumping scenarios may be constructed by the pumping well manager, as is shown, for example, by Brak-

enhoff et al. (2022). In the latter case, well managers may explore the combined effect of meteorological stresses and different pumping regimes.

For natural stresses (e.g., river levels), the forecasted input data may be the output of another impact model (e.g., a rainfall-runoff model), and one possibility is to test chaining different impact models (i.e., force the groundwater model with additional ensembles of river levels). Alternatively, machine and deep learning approaches may be tried to directly include the effects of upstream precipitation on the groundwater levels in downstream aquifers. Deep learning models have also been found useful for generating ensemble predictions of water temperature in Switzerland (Padrón et al., 2025), which is also influenced by precipitation higher up in the catchment, and could also be tested to directly provide groundwater level predictions. Another improvement and avenue for future research is using deep learning techniques in a hybrid approach, for example, to post-process the forecast from the lumped groundwater model (Slater et al., 2023).

The prediction intervals were found to be both under- and overconfident in this study. More research is needed to improve the prediction intervals. Using more complex noise models that better describe the error structure may seem obvious, but may be difficult to apply when dealing with the irregular time steps between groundwater level measurements often found in real-world groundwater level data. Another (possibly more practical) approach is to use alternative, non-parametric methods for estimating prediction intervals, such as bootstrapping or quantile-based prediction intervals. Future research can investigate if such alternatives provide better prediction intervals.

A final challenge is the upscaling of the groundwater forecast system to provide more spatially distributed information. An obvious first step would be to increase the number of monitoring wells when they are available to improve the spatial coverage. Because it can be challenging to obtain the groundwater level data in an operational setting for many wells, an alternative approach would be to use representative wells. In this approach, wells are chosen to represent other wells based on a study of the correlations between the available time series (see, for example, Wunsch et al., 2022, and the GRUVO platform https://gruvo.bgr.de/, last accessed 2026-08-27). When the absolute values are not of interest and post-processing is not viable, an entirely different approach would be to develop a spatially distributed groundwater model. In that case, a surrogate model may be required to solve the computational problem of running large ensembles to generate the forecasts. More research is needed to investigate the feasibility of such approaches.

6 Summary and conclusion

In this study, an ensemble groundwater system was developed to forecast groundwater levels 32 days ahead in Switzerland, using meteorological ensemble predictions and lumped-parameter groundwater models. The system was tested and evaluated for 12 years using historic forecast data (2012-2023) and measured groundwater levels for the same period. Data from six monitoring wells in alluvial aquifers in Switzerland were used to test the quality and skill of the forecasts generated by the system. These wells represent diverse hydrogeological and climatic settings, differing in altitude, groundwater dynamics, and response times to recharge processes driven by precipitation and snowmelt.

Based on the results, it is concluded that skillful forecasts can be made for the investigated lead times (32 days) using the proposed ensemble groundwater prediction (EGP) system. The required level of data-model complexity to produce reliable forecasts depends on the memory time of the aquifer system, which serves as an indicator of its reactivity and dynamics in response to infiltration. The results suggest that, for the investigated monitoring wells and lead times, greater model complexity is justified for short-memory systems that respond rapidly to recharge, whereas simpler models may suffice for long-memory systems with slower, more buffered responses. For long-memory systems, the use of a lumped groundwater model and climatology suffices to produce good forecasts at a 32-day lead time. While these results may be expected, the findings from this study confirm this expectation and provide valuable information for the design and development of future operational groundwater level prediction systems. From design the settings presented here are fully compatible to be added to the research platform on drought impact "drought.ch" (Padrón et al., 2026).

Future research should investigate if the ensemble groundwater prediction system can also be applied to produce forecasts with larger lead times, up to a few months, an important time frame for groundwater resources management. Forecasting groundwater levels in mountainous regions, such as Davos, remains challenging and might require more advanced bias-correction and post-processing techniques to produce skillful forecasts.

Code and data availability. All data and code necessary to reproduce the figures and tables from this study are available from the following Zenodo repository: https://doi.org/10.5281/zenodo.17171933

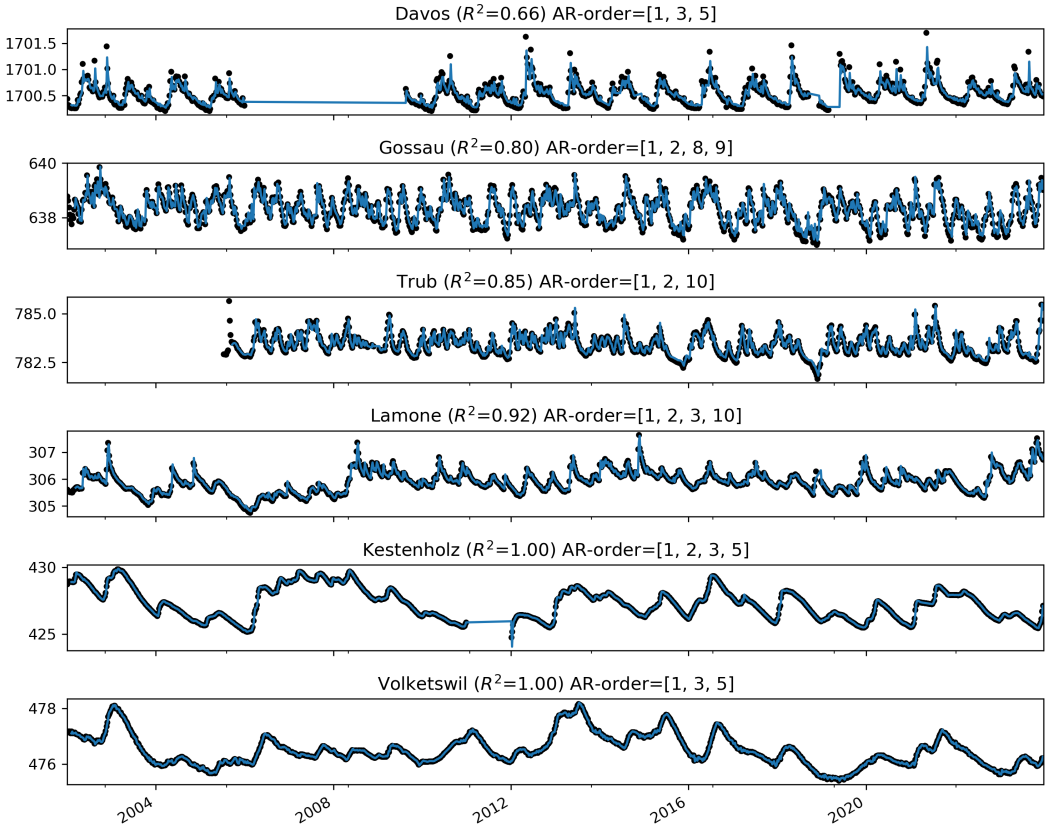

Figure A1. Measured and simulated groundwater levels from the autoregressive models fitted to the entire observation period. The estimated AR model order for each monitoring well is given above each plot, along with the R^2 computed over the entire period. Note that the metrics are rounded off to two decimals and never exactly 1.0.

Appendix A: Autoregressive models

Autoregressive (AR) models were constructed for each monitoring well to assess how well a persistence-based forecast might predict groundwater levels. The models were developed with the Python package Statsmodels (Seabold and Perktold, 2010) using weekly groundwater level measurements. The model order was automatically selected using the Bayesian Information Criterion (BIC) from all possible AR combinations up to order 10 (i.e., using groundwater level fluctuations from the last 10 weeks). Figure A1 shows the resulting groundwater level simulations for each well. As the figure shows, model performance increases with increasing memory time, but the AR models generally simulate groundwater levels with high accuracy.

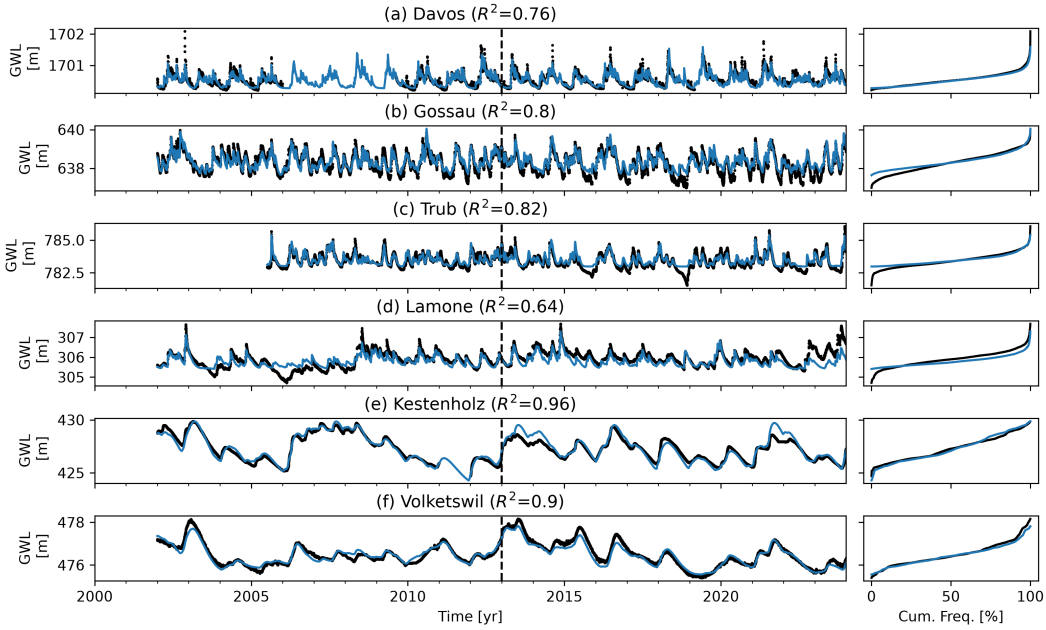

Figure B1. Measured and simulated groundwater levels for the initial Pastas models, calibrated to the data from 2002-2012. The vertical dashed line denotes the end of the calibration period and the start of the validation period. The reported R^2 in the figure is computed over the entire period 2002-2023 and therefore does not match the values in Table 2. The plots on the right show the cumulative frequency distribution.

Appendix B: Simulated groundwater levels and autocorrelation plots

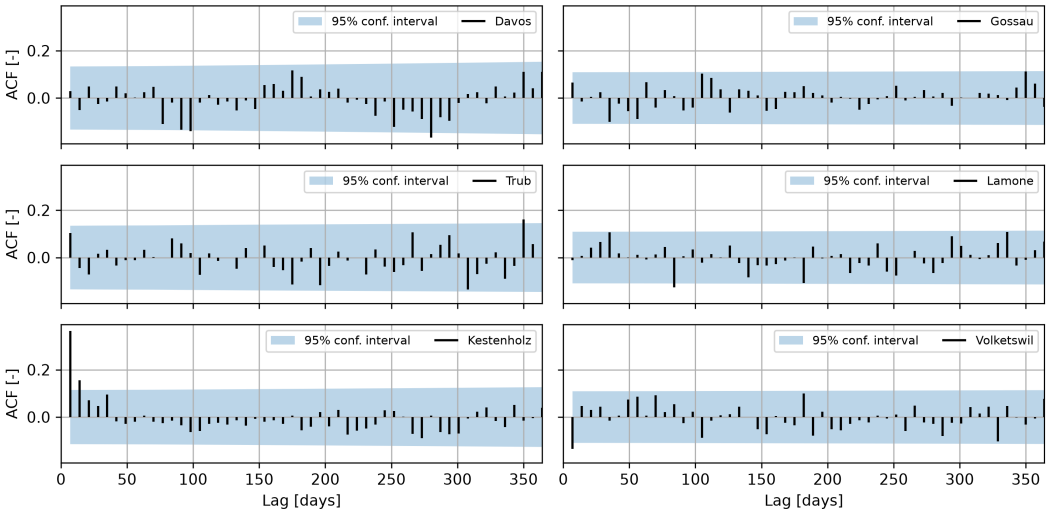

Figure B2. Plots of the autocorrelation functions for the initial Pastas models calibrated to the data from 2002 to 2012. The black bars denote the autocorrelation in the noise at various time lags, and the blue shaded areas denote the 95% confidence intervals.

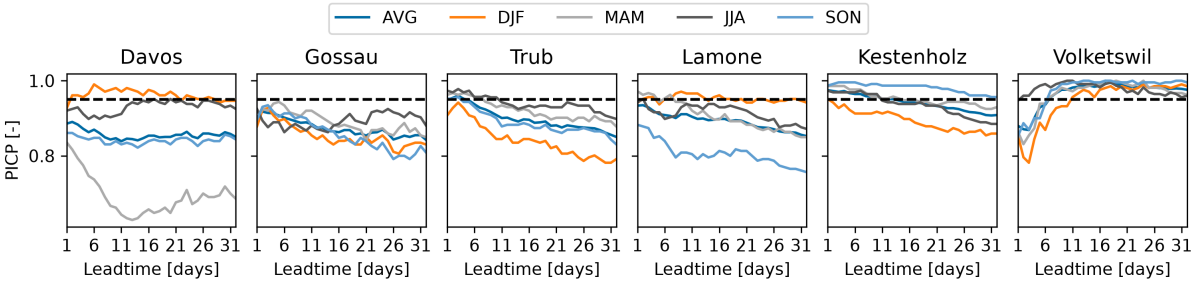

Figure C1. Values of the prediction interval coverage probability (PICP) for each monitoring well (each subplot) are split by the seasons (see the caption of Table 2 for the coding of the seasons).

Appendix C: Seasonal PICP values

Author contributions. RAC performed the formal analysis with help of KB. RAC wrote the original draft with comments and edits from KB, MZ, CM, and MS. MZ curated the meteorological forecasts. RAC conceptualized the study with the help of KB and MZ. MS obtained the funding for this research.

Competing interests. The Authors declare that no competing interests are present.

Acknowledgements. This research was funded by Eawag Discretionary funds through the project "Groundwater extremes". The contribution of MZ and KB has been supported by the Malefix project, which is funded by WSL as part of the WSL Program Extremes.

References

BAFU (2022). Gewässer in der Schweiz. Technical Report Umwelt-Zustand Nr. 2207, Bundesamt für Umwelt BAFU.

Bauer, V. M. and Scherrer, S. C. (2024). The observed evolution of sub-daily to multi-day heavy precipitation in Switzerland. *Atmospheric Science Letters*, 25(9):e1240. Publisher: John Wiley & Sons, Ltd.

Blitzstein, J. K. and Hwang, J. (2019). *Introduction to Probability*. CRC Press, Boca Raton, FL, 2nd edition.

Bogner, K., Chang, A. Y.-Y., Bernhard, L., Zappa, M., Monhart, S., and Spirig, C. (2022). Tercile Forecasts for Extending the Horizon of Skillful Hydrological Predictions. *Journal of Hydrometeorology*, 23(4):521 – 539. Place: Boston MA, USA Publisher: American Meteorological Society.

Bogner, K., Liechti, K., Bernhard, L., Monhart, S., and Zappa, M. (2018). Skill of Hydrological Extended Range Forecasts for Water Resources Management in Switzerland. *Water Resources Management*, 32(3):969–984.

Bordes, M., Luis, J., Gurdak, J. J., and others (2011). Climate change effects on groundwater resources a global synthesis of findings and recommendations. Publisher: Leiden, Netherlands CRC Press Balkema Book.

Brakenhoff, D. A., Vonk, M. A., Collenteur, R. A., Van Baar, M., and Bakker, M. (2022). Application of Time Series Analysis to Estimate Drawdown From Multiple Well Fields. *Frontiers in Earth Science*, 10. location=The Netherlands.

Brunner, M. I., Liechti, K., and Zappa, M. (2019). Extremeness of recent drought events in Switzerland: dependence on variable and return period choice. *Natural Hazards and Earth System Sciences*, 19(10):2311–2323.

Cantone, C., Ivars Grape, H., El Habash, S., and Pechlivanidis, I. G. (2023). A co-generation success story: Improving drinking water management through hydro-climate services. *Climate Services*, 31:100399.

Collenteur, R. A., Bakker, M., Caljé, R., Klop, S. A., and Schaars, F. (2019). Pastas: Open Source Software for the Analysis of Groundwater Time Series. *Groundwater*, 57(6):877–885.

Collenteur, R. A., Bakker, M., Klammler, G., and Birk, S. (2021). Estimation of groundwater recharge from groundwater levels using nonlinear transfer function noise models and comparison to lysimeter data. *Hydrology and Earth System Sciences*, 25(5):2931–2949. location=Austria.

Collenteur, R. A., Moeck, C., Schirmer, M., and Birk, S. (2023). Analysis of nationwide groundwater monitoring networks using lumped-parameter models. *Journal of Hydrology*, 626:130120.

Duan, Q., Pappenberger, F., Wood, A., Cloke, H. L., and Schaake, J. C. (2019). *Handbook of hydrometeorological ensemble forecasting*, volume 10. Springer Berlin/Heidelberg, Germany.

Epting, J., Michel, A., Affolter, A., and Huggenberger, P. (2021). Climate change effects on groundwater recharge and temperatures in Swiss alluvial aquifers. *Journal of Hydrology X*, 11:100071.

Gneiting, T., Raftery, A. E., Westveld, A. H., and Goldman, T. (2005). Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation. *Monthly Weather Review*, 133(5):1098–1118. Place: Boston MA, USA Publisher: American Meteorological Society.

Halloran, L. J., Millwater, J., Hunkeler, D., and Arnoux, M. (2023). Climate change impacts on groundwater discharge-dependent streamflow in an alpine headwater catchment. *Science of The Total Environment*, 902:166009.

Hersbach, H. (2000). Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. *Weather and Forecasting*, 15(5):559–570. Place: Boston MA, USA Publisher: American Meteorological Society.

Ho, C. K., Hawkins, E., Shaffrey, L., Bröcker, J., Hermanson, L., Murphy, J. M., Smith, D. M., and Eade, R. (2013). Examining reliability of seasonal to decadal sea surface temperature forecasts: The role of ensemble dispersion. *Geophysical Research Letters*, 40(21):5770–5775. Publisher: John Wiley & Sons, Ltd.

Huang, J.-Y. and Shih, D.-S. (2020). Assessing Groundwater Level with a Unified Seasonal Outlook and Hydrological Modeling Projection. *Applied Sciences*, 10(24).

Jolliffe, I. T. and Stephenson, D. B. (2012). *Forecast verification: a practitioner's guide in atmospheric science*. John Wiley & Sons.

Mackay, J., Jackson, C., Brookshaw, A., Scaife, A., Cook, J., and Ward, R. (2015). Seasonal forecasting of groundwater levels in principal aquifers of the United Kingdom. *Journal of Hydrology*, 530:815–828.

MeteoSwiss (2022). RhiresD and TabsD gridded precipitation and temperature data sets.

Moeck, C., Brunner, P., and Hunkeler, D. (2016). The influence of model structure on groundwater recharge rates in climate-change impact studies. *Hydrogeology Journal*, 24(5):1171–1184.

Monhart, S., Spirig, C., Bhend, J., Bogner, K., Schär, C., and Liniger, M. A. (2018). Skill of Subseasonal Forecasts in Europe: Effect of Bias Correction and Downscaling Using Surface Observations. *Journal of Geophysical Research: Atmospheres*, 123(15):7999–8016. Publisher: John Wiley & Sons, Ltd.

Monhart, S., Zappa, M., Spirig, C., Schär, C., and Bogner, K. (2019). Subseasonal hydrometeorological ensemble predictions in small- and medium-sized mountainous catchments: benefits of the NWP approach. *Hydrology and Earth System Sciences*, 23(1):493–513.

Neumann, J. L., Arnal, L., Emerton, R. E., Griffith, H., Hyslop, S., Theofanidi, S., and Cloke, H. L. (2018). Can seasonal hydrological forecasts inform local decisions and actions? A decision-making activity. *Geoscience Communication*, 1(1):35–57.

Padrón, R. S., Buchecker, M., Cremona, A., Farinotti, D., Gossner, M. M., Maidl, E., Pellissier, L., Pezzatti, G. B., Minetti, A., Bernhard, L., Zappa, M., and Bogner, K. (2026). An operational platform for sub-seasonal forecasts of drought-related hazards. *International Journal of Cartography*, 0(0):1–22.

Padrón, R. S., Zappa, M., Bernhard, L., and Bogner, K. (2025). Extended-range forecasting of stream water temperature with deep-learning models. *Hydrology and Earth System Sciences*, 29(6):1685–1702.

Prudhomme, C., Hannaford, J., Harrigan, S., Boorman, D., Knight, J., Bell, V., Jackson, C., Svensson, C., Parry, S., Bachiller-Jareno, N., Davies, H., Davis, R., Mackay, J., McKenzie, A., Rudd, A., Smith, K., Bloomfield, J., Ward, R., and Jenkins, A. (2017). Hydrological Outlook UK: an operational streamflow and groundwater level forecasting system at monthly to seasonal time scales. *Hydrological Sciences Journal*, 62(16):2753–2768. Publisher: Taylor & Francis.

Robertson, D. E., Fu, G., Barron, O., Hodgson, G., and Schepen, A. (2024). A new approach of coupled long-range forecasts for streamflow and groundwater level. *Journal of Hydrology*, 631:130837.

Rodell, M., Famiglietti, J. S., Wiese, D. N., Reager, J. T., Beaudoing, H. K., Landerer, F. W., and Lo, M.-H. (2018). Emerging trends in global freshwater availability. *Nature*, 557(7707):651–659.

Scanlon, B. R., Fakhreddine, S., Rateb, A., de Graaf, I., Famiglietti, J., Gleeson, T., Grafton, R. Q., Jobbagy, E., Kebede, S., Kolusu, S. R., Konikow, L. F., Long, D., Mekonnen, M., Schmied, H. M., Mukherjee, A., MacDonald, A., Reedy, R. C., Shamsudduha, M., Simmons, C. T., Sun, A., Taylor, R. G., Villholth, K. G., Vörösmarty, C. J., and Zheng, C. (2023). Global water resources and the role of groundwater in a resilient water future. *Nature Reviews Earth & Environment*, 4(2):87–101.

Scherrer, S. C., Hirschi, M., Spirig, C., Maurer, F., and Kotlarski, S. (2022). Trends and drivers of recent summer drying in Switzerland. *Environmental Research Communications*, 4(2):025004. Publisher: IOP Publishing.

Seabold, S. and Perktold, J. (2010). Statsmodels: Econometric and Statistical Modeling with Python. In *Proceedings of the 9th Python in Science Conference*, volume 57, page 61.

Slater, L. J., Arnal, L., Boucher, M.-A., Chang, A. Y.-Y., Moulds, S., Murphy, C., Nearing, G., Shalev, G., Shen, C., Speight, L., Villarini, G., Wilby, R. L., Wood, A., and Zappa, M. (2023). Hybrid forecasting: blending climate predictions with AI models. *Hydrology and Earth System Sciences*, 27(9):1865–1889.

Tuel, A., Schaefli, B., Zscheischler, J., and Martius, O. (2022). On the links between sub-seasonal clustering of extreme precipitation and high discharge in Switzerland and Europe. *Hydrology and Earth System Sciences*, 26(10):2649–2669.

Van der Spek, J. E. and Bakker, M. (2017). The influence of the length of the calibration period and observation frequency on predictive uncertainty in time series modeling of groundwater dynamics. *Water Resources Research*, 53(3):2294–2311.

Viel, C., Beaulant, A.-L., Soubeyroux, J.-M., and Céron, J.-P. (2016). How seasonal forecast could help a decision maker: an example of climate service for water resource management. *Advances in Science and Research*, 13:51–55.

von Asmuth, J. R. and Bierkens, M. F. P. (2005). Modeling irregularly spaced residual series as a continuous stochastic process. *Water Resources Research*, 41(12):W12404.

von Asmuth, J. R., Bierkens, M. F. P., and Maas, K. (2002). Transfer function-noise modeling in continuous time using predefined impulse response functions. *Water Resources Research*, 38(12):23–1–23–12.

von Asmuth, J. R., Maas, K., Bakker, M., and Petersen, J. (2008). Modeling Time Series of Ground Water Head Fluctuations Subjected to Multiple Stresses. *Groundwater*, 46(1):30–40.

Vremec, M., Collenteur, R. A., and Birk, S. (2024). *PyEt* v1.3.1: a Python package for the estimation of potential evapotranspiration. *Geoscientific Model Development*, 17(18):7083–7103.

Wahba, G. (1990). *Spline models for observational data*. SIAM.

Wanders, N., Thober, S., Kumar, R., Pan, M., Sheffield, J., Samaniego, L., and Wood, E. F. (2019). Development and Evaluation of a Pan-European Multimodel Seasonal Hydrological Forecasting System. *Journal of Hydrometeorology*, 20(1):99–115.

White, C. J., Carlsen, H., Robertson, A. W., Klein, R. J., Lazo, J. K., Kumar, A., Vitart, F., Coughlan de Perez, E., Ray, A. J., Murray, V., Bharwani, S., MacLeod, D., James, R., Fleming, L., Morse, A. P., Eggen, B., Graham, R., Kjellström, E., Becker, E., Pegion, K. V., Holbrook, N. J., McEvoy, D., Depledge, M., Perkins-Kirkpatrick, S., Brown, T. J., Street, R., Jones, L., Remenyi, T. A., Hodgson-Johnston, I., Buontempo, C., Lamb, R., Meinke, H., Arheimer, B., and Zebiak, S. E. (2017). Potential applications of subseasonal-to-seasonal (S2S) predictions. *Meteorological Applications*, 24(3):315–325. Publisher: John Wiley & Sons, Ltd.

Wunsch, A., Liesch, T., and Broda, S. (2022). Feature-based Groundwater Hydrograph Clustering Using Unsupervised Self-Organizing Map-Ensembles. *Water Resources Management*, 36(1):39–54.