

Forecasting groundwater levels at unseen wells without per-well training: GeoAI-GWFM-KR, a time-series foundation model adapted to the Korean monitoring network

Technical note, GeoAI-GWFM-KR v1.0

Eungyu Park, Jangwon Park and Taeyu Kim (GeoAI Alignment, Inc., Daegu, Republic of Korea)

October 2026

Key points

- A time-series foundation model adapted once on 188 Korean monitoring wells forecasts wells at unseen stations without per-well training
- With the same input variables it was more accurate than long short-term memory networks (LSTMs) trained on each well at 39 of 46 wells at 30 days
- Without any adaptation the pretrained model showed no significant difference from per-well LSTMs at 30 and 90 days

Abstract

Groundwater-level forecasts at monitoring wells are still built mostly by training a recurrent network on each well's own record, so every new well needs its own model and a well with a short record gets a weak one. Pretrained time-series foundation models forecast series they have never seen, which raises a practical question: does training a network for each well still buy accuracy? We tested this at 30 monitoring stations (46 wells) of the Korean National Groundwater Monitoring Network that were held out before any model was trained, with weekly forecasts for 2023–2024 and every forecaster given the same input variables, the observed levels and the daily rainfall. Chronos-2 used without any training showed no significant difference from per-well long short-term memory (LSTM) networks at both 30 and 90 days, and was less accurate than them at 7 days. Adapted once with a low-rank adapter trained on 188 wells at other stations, the model, released as GeoAI-GWFM-KR, reached a median skill against persistence of 0.41, 0.48 and 0.52 at 7, 30 and 90 days, against 0.37, 0.30 and 0.33 for the per-well LSTM, and was the more accurate at 34, 39 and 40 of the 46 wells; at wells installed since 2019 it was the more accurate at 13 of 13 at 30 days. A physics-informed loss did not make the per-well networks more accurate. For 7- to 90-day forecasts with a rainfall forecast at wells with at least a year of record, one adapted pretrained model matched or exceeded per-well training, without any training at the new well.

1. Introduction

A water agency that installs a monitoring well today will want a forecast of its level for the coming weeks long before the well has a record worth learning from. The usual route is to collect a few years of data and then train a recurrent network, a long short-term memory (LSTM) network or a gated recurrent unit (GRU), on that well's record, often with rainfall as an input and sometimes with a physical relation added to the loss. The route repeats for every well. It is costly, it leaves new wells without a model for years,

and its result is seldom compared with a simple reference such as persistence or with a model that never saw the well.

Data-driven groundwater forecasting has largely followed this route. Recurrent networks (Hochreiter & Schmidhuber, 1997; Cho et al., 2014) have been compared with convolutional and autoregressive networks trained for individual wells (Wunsch et al., 2021), and convolutional networks trained one per well were used for projections at more than one hundred wells (Wunsch et al., 2022). In surface hydrology, LSTMs proved strong rainfall–runoff models (Kratzert et al., 2018), and a single LSTM trained on hundreds of catchments with catchment attributes matched or exceeded models calibrated to each catchment (Kratzert et al., 2019). For groundwater, such global models have been harder to build: the static features meant to distinguish wells did not let a model trained on many wells generalize to new ones (Heudorfer et al., 2024). Another line adds physical relations to the training loss to help networks with scarce data (Read et al., 2019), and what role such knowledge keeps once data-driven models predict well is an open question (Nearing et al., 2021).

Time-series foundation models change the starting point. Models such as TimesFM (Das et al., 2024), Moirai (Woo et al., 2024), MOMENT (Goswami et al., 2024) and Chronos (Ansari et al., 2024, 2025) are pretrained on large and varied collections of series and forecast a new series from its recent history without task-specific training; some, such as Chronos-2, also take known covariates. On broad benchmarks such as GIFT-Eval (Aksu et al., 2024) they are competitive with models trained for each dataset. In ten-day forecasts at three Korean monitoring wells with twenty-year records, TimesFM-3 used without training was level with or ahead of tuned recurrent networks trained on ten years of record, most of the short-horizon skill came from the weather of the forecast days, and Chronos-2 gained little from that weather (Kim et al., 2026). Whether a pretrained model, used as it is or adapted once, matches tuned per-well networks at horizons of one to three months, across many wells and at stations it has never seen, has not been tested.

Here we ask three questions. Does a pretrained model used without any training forecast unseen wells as accurately as recurrent networks trained on each well? Does adapting the pretrained model once, on other wells of the same network, change the answer? And where does the advantage of each approach lie? We answer them with GeoAI-GWFM-KR, the Chronos-2 model (Ansari et al., 2025) with a low-rank adapter trained on the Korean National Groundwater Monitoring Network, in a benchmark built so that it favours neither side: the same input variables, identical training and validation years, test stations kept out of all training, and identical scoring days.

2. Data and benchmark

Wells and periods. We used daily groundwater levels of 340 wells at 216 stations of the National Groundwater Monitoring Network (National Groundwater Information Center, 2026), 148 of the eligible stations in the Yeongnam region (Gyeongsangbuk-do, Gyeongsangnam-do, Daegu, Busan and Ulsan) and 5 in Gangwon-do and Jeollabuk-do, with daily rainfall from a paired weather station (Korea Meteorological Administration, 2026), 2005–2024. Most stations have a bedrock well, and many also have an alluvial or shallow well. Training data end on 31 December 2021, 2022 is the validation year for every model choice, and 2023–2024 is the test period. A well was eligible with at least two years of valid levels before 2023 and at least one year in 2023–2024; 234 wells at 153 stations qualified.

Test stations. Before any model was run, we held out 20% of the stations, stratified by province with a fixed random seed. The 46 wells at these 30 stations (27 bedrock, 18 alluvial, 1 unclassified) form the test set; 29 of these stations are in the Yeongnam region and 1 in Gangwon-do. The 188 wells at the other 123 stations are used for training. No test well shares a station with a training well. Of the test wells, 28 have records starting in 2012 or earlier and 13 were installed in 2019 or later.

Inputs and scoring. Every forecaster receives the same variables, the observed levels up to the day before the forecast and the daily rainfall, in the form its architecture needs (Table 1): Chronos-2 takes the raw series with its gaps and up to 1,095 days of history, and the recurrent networks take the level with the last observation carried forward over gaps, an observed flag and rainfall sums derived from the rainfall, over 120 or 365 days. Only gross outliers are removed from the inputs (values more than ten interquartile ranges from the median, both taken from data up to 2021; 0.41% of observations), so no input depends on later data. Scoring uses observed days that pass a quality check on the whole record: runs of ten or more identical values and isolated one-day spikes (a Hampel-type filter (Hampel, 1974) that flags deviations from the seven-day running median larger than six times the running median absolute deviation, keeping rises within three days of at least 10 mm of rain) are excluded from scoring, 0.61% of observations, and stay in the inputs.

Forecasting task. From each forecast origin, every seventh day of the test period, each model forecasts the daily level 7, 30 and 90 days ahead. In the main setting the observed rainfall of the forecast days is given, which stands in for a perfect rainfall forecast; in a second setting no rainfall is given for the forecast days. Accuracy is the root-mean-square error (RMSE) over the scored forecast days of a well, and skill is $1 - \text{RMSE} / \text{RMSE of persistence}$, which holds the last observed level. This is the skill score with persistence as the reference (Murphy, 1988), written with the RMSE; it is more conservative than the usual mean-square-error form, $1 - (\text{RMSE} / \text{RMSE of persistence})^2$, and puts wells with small and large level variations on one scale. Because persistence degrades with lead time, skill compares forecasters at the same horizon, not across horizons. The score for a horizon pools all lead days from 1 to H; scoring only day H gives the same ranking (Section 4.2). All forecasters are scored on the same days. Comparisons are paired by well with the Wilcoxon signed-rank test (Wilcoxon, 1945); because wells at one station are not independent, we repeated each test on station means of the paired differences.

3. Forecasters

Chronos-2 and the adapter. Chronos-2 is a 120-million-parameter encoder model that forecasts quantiles of a target series from its history and from past and known-future covariates (Ansari et al., 2025). Its pretraining corpus consists of public forecasting benchmark collections and synthetic series; it was not trained on the groundwater records used here. We give it up to 1,095 days of history and daily rainfall as a covariate, and use the median as the point forecast. GeoAI-GWFM-KR adds low-rank adaptation (LoRA; Hu et al., 2022) matrices of rank 8 to the attention projections and the output layer, 1.2 million trainable parameters or 1.0% of the model. Each training well enters training twice, with rainfall known for the forecast days and with rainfall as a past-only covariate, so one adapter serves both settings. Windows are drawn at random positions in the training records up to 2021 with a 90-day forecast length and the Chronos-2 quantile loss. Four candidate settings (learning rate 1×10^{-4} or 3×10^{-4} , 1,000 or 3,000 steps) were scored on the training wells in 2022, and a rule fixed in advance took the highest mean median skill over the three horizons and two rainfall settings (selected: 1×10^{-4} and 1,000 steps,

validation score 0.279 against 0.21–0.27 for the others and 0.182 without the adapter). The selected setting was trained with three seeds; results average them, and the released adapter is seed 0, designated before testing.

Per-well recurrent networks. For every test well we trained single-layer networks on that well’s own record, following the recursive design of Park et al. (2026b). The network reads a window of past days, each described by the level (the last observation carried forward over gaps), its daily change, an indicator of whether the day was observed, the daily rainfall and its 30-, 90- and 365-day sums. It then advances one day at a time: it predicts the daily change from its state and that day’s rainfall, adds it to the level and feeds the predicted level back, for as many days as the forecast horizon. Training uses the same rollout length as the forecast, a loss on observed days only, the Adam optimizer, and early stopping on 2022. The per-well LSTM chose its hidden size (64 or 128 units) and window (120 or 365 days) separately for each well and horizon by the three-seed validation loss. A GRU (64 units, 120 days) and an LSTM with the groundwater fluctuation equation of Park and Parker (2008) as an additional loss (64 units, 120 days, loss weight chosen per well from 0.1, 1 and 10) complete the per-well set. The selected networks passed a competence check (training lead-1 error below persistence) at 42 or more of the 46 wells. These networks see the test well’s own past, which the adapter never does.

Regional LSTM. To separate the effect of pooling many wells from the effect of pretraining, one LSTM was trained on all 188 training wells with the inputs of the per-well networks, standardized per well with each well’s data up to 2021, and applied to the test stations. It uses exactly the training data of the adapter. It chose among six settings by validation loss: hidden size 64 or 128 with a window of 120 or 365 days at a learning rate of 10^{-3} , and hidden size 64 or 128 with a 120-day window at 10^{-4} . The two settings at 10^{-4} were added after first regional runs at 10^{-3} , made under an earlier input design, had reached their best validation loss in their first epochs and had been scored on the test period; the final choice among all six used the 2022 validation loss only (selected at every horizon: hidden size 64, window 120 days, learning rate 0.0001).

Design for a neutral comparison. Table 1 lists what each forecaster receives, and Table A1 gives the training settings. All forecasters see the same variables of a well, the observed levels and the past and forecast-day rainfall; none receives calendar features, weather other than rainfall, or training data after 2021. Input forms and history lengths follow each architecture. The per-well networks have one advantage that the adapter does not: they learn from the test well’s own record. Every trained model chose its settings on 2022 and was trained with three seeds, and all forecasters are scored on the same days. The candidate sets differ in size; to check that the tuning of the per-well LSTM does not decide the comparison, we also chose each well’s LSTM setting with hindsight on the test period (Section 4.2).

Table 1. Information, training data and tuning of each forecaster. All settings were chosen on the validation year 2022 by the criterion given; test stations were never used.

Forecaster	Learns from	Inputs	Candidate settings	Selection	Seeds
GeoAI-GWFM-KR	188 training wells, to 2021	levels with gaps, up to 1,095 days; past and forecast-day rainfall	4 (learning rate $\times$ steps)	median skill on training wells, 2022	3

Forecaster	Learns from	Inputs	Candidate settings	Selection	Seeds
Chronos-2	nothing	as GeoAI-GWFM-KR	none	none	deterministic
Per-well LSTM	the test well, to 2021	level carried forward over gaps with an observed flag, 120 or 365 days; rainfall and its 30-, 90-, 365-day sums	4 (hidden size $\times$ window), per well and horizon	validation loss, 3-seed mean	3
Per-well GRU	the test well, to 2021	as per-well LSTM, 120 days	1	early stopping only	3
Per-well LSTM + physics loss	the test well, to 2021	as per-well LSTM, 120 days	3 (loss weight), per well and horizon	validation loss, 3-seed mean	3
Regional LSTM	188 training wells, to 2021	as per-well LSTM, 120 or 365 days	6 (hidden size $\times$ window $\times$ learning rate)	validation loss, 3-seed mean	3

4. Results

4.1 A pretrained model without training

Chronos-2 used without any training reached a median skill of 0.21, 0.26 and 0.31 at 7, 30 and 90 days (Table 2, Figure 1; mean errors in Table A2). Against the per-well LSTM, trained on each test well’s own record, it was the more accurate at 7, 17 and 16 of the 46 wells ($p < 0.001$, $p = 0.084$ and $p = 0.49$; station-level $p = 0.0011$, $p = 0.1$ and $p = 0.43$; all paired tests in Table A3). It thus showed no significant difference from the per-well LSTM at both 30 and 90 days, although the point estimates favoured the LSTM, and was less accurate than it at 7 days; scored on the last day of each horizon only, the 30- and 90-day comparisons were again not significant ($p = 0.84$ and $p = 0.51$). Against the per-well GRU, the strongest per-well network at 30 and 90 days, it was less accurate than it at 7 days, was less accurate than it at 30 days and showed no significant difference from it at 90 days. Against the regional LSTM, trained on the same 188 wells that the adapter uses, it was the more accurate at 20, 29 and 30 wells ($p = 0.63$, $p = 0.023$ and $p = 0.002$).

Table 2. Median skill against persistence on the 46 test wells, 2023–2024, with the rainfall of the forecast days given. Trained models are averaged over three seeds.

Forecaster	Training data	7 days	30 days	90 days
GeoAI-GWFM-KR	188 training wells (adapter)	0.41	0.48	0.52
Chronos-2	none	0.21	0.26	0.31
Per-well LSTM	the test well’s own record	0.37	0.30	0.33
Per-well GRU	the test well’s own record	0.41	0.33	0.37

Forecaster	Training data	7 days	30 days	90 days
Per-well LSTM with physics loss	the test well's own record	0.39	0.29	0.34
Regional LSTM	188 training wells	0.25	0.20	0.24

4.2 Adapted once, applied to unseen stations

The adapter raised the median skill to 0.41, 0.48 and 0.52, and GeoAI-GWFM-KR was more accurate than Chronos-2 alone at 44, 45 and 46 of the 46 wells (Figure 2a); the three training seeds gave 0.41–0.42, 0.48–0.49 and 0.50–0.53. Against the per-well LSTM it was the more accurate at 34, 39 and 40 wells ($p = 0.0014$, $p < 0.001$ and $p < 0.001$; station-level $p = 0.0062$, $p < 0.001$ and $p < 0.001$). Against the per-well GRU the counts were 32, 38 and 35, against the physics-guided LSTM 31, 42 and 43, and against the regional LSTM 42, 46 and 44. Two checks show that the comparison does not hinge on how forecasts were scored or how the per-well LSTM was tuned. Scored on the last day of each horizon only, GeoAI-GWFM-KR was more accurate than the per-well LSTM at 33, 42 and 36 wells ($p \leq 0.001$). With each well's LSTM setting chosen with hindsight on the test period, the most favourable choice for the baseline, it remained the more accurate at 37 and 37 wells at 30 and 90 days ($p < 0.001$ and $p < 0.001$), while at 7 days the difference was not significant (25 of 46, $p = 0.63$). Figure 3 shows the forecasts at the test well with the median gain over the per-well LSTM.

4.3 Where per-well training still competes

The advantage depends on how much record a per-well network can learn from (Figure 4). At the 13 wells installed in 2019 or later, with one to three years of training record, GeoAI-GWFM-KR was the more accurate at 12, 13 and 13 wells at 7, 30 and 90 days. At the 28 wells with records from 2012 or earlier, a per-well network trained on ten or more years is a strong forecaster: there the counts were 19, 21 and 23 of 28 ($p = 0.13$, $p < 0.001$ and $p < 0.001$).

4.4 What the recurrent design choices added

Two choices that are often made to improve recurrent forecasters did not help here. Adding the fluctuation equation as a loss did not change the accuracy of the per-well LSTM with the same hidden size and window ($p \geq 0.35$ at all horizons), in line with our study of the same equation as a training loss, where it stopped adding information once a network had several years of record (Park et al., 2026b). Pooling the training wells into one regional LSTM did not substitute for the well's own record: the regional LSTM was less accurate than the per-well LSTM at 37, 33 and 33 of the 46 wells. The pooled data helped only when they adapted a pretrained model.

4.5 Response to rainfall

To see how the forecasts respond to rainfall we added 100 mm of rain on the third forecast day, at origins every 28 days in 2023–2024. The median 30-day forecast of GeoAI-GWFM-KR rose by 0.35 m and peaked 2 days after the storm; Chronos-2 alone rose by 0.05 m and peaked after 5 days. In the same wells' records before 2023, the level rose by a median of 0.57 m (interquartile range 0.34–0.96 m) after a day with at least 80 mm of rain, peaking 2 days later; these events often include rain on neighbouring days, so only the order of magnitude is comparable. Scaling the forecast-day rainfall by 0, 1 and 2 ordered the 30-day levels in the expected sense in 99.5% of the adapter's forecasts and 93.7% of those of Chronos-2 alone. Without rainfall for the forecast days, the median skill of the adapter was 0.07, 0.16 and

0.19 against 0.05, 0.11 and 0.18 for Chronos-2 alone, a smaller gain that was not significant ($p = 0.2$) at 7 days and significant ($p < 0.001$) at 30 days and significant ($p < 0.001$) at 90 days.

4.6 Training-set size and forecast intervals

Adapters trained on random subsets of the training wells (seed 0) reached a 30-day median skill of 0.43 with 10 wells, 0.46 with 40, 0.48 with 100 and 0.48 with all 188, against 0.26 without adaptation (Figure 2b). The nominal 80% interval (10–90% quantiles) of the released adapter covered 80%, 77% and 74% of observations at 7, 30 and 90 days with forecast-day rainfall, and 79%, 72% and 70% without it; Chronos-2 alone covered 76–82%.

5. Discussion

5.1 What per-well training buys in the foundation-model era

The case for training a recurrent network at every monitoring well has rested on accuracy: a model that has learned the well should forecast it best. In this benchmark, for forecasts of 7 to 90 days at wells with at least a year of record, per-well training did not buy that accuracy. A pretrained model that had never seen these records showed no significant difference from the per-well LSTM at both 30 and 90 days, although its point estimates favoured the LSTM, and after one adaptation on other stations it was the more accurate at most wells at every horizon, with the largest margins at the wells that per-well training serves worst, those with short records. Per-well training competed best at 7 days and at wells with a decade or more of record: there the adapted model showed no significant difference from the per-well LSTM at 7 days, and with each well's LSTM tuned in hindsight the 7-day difference disappeared. The two refinements most often added to per-well networks, a physics-informed loss and pooling across wells, did not close the gap; the weak transfer of the pooled LSTM to new stations matches the difficulty of global groundwater models reported by Heudorfer et al. (2024).

The practical consequence is a change of default for this task. For a recently installed well with a year or more of record, a forecast no longer waits for a training campaign; it comes from one model shared by the network. Wells with less than a year of record were not tested. Effort that used to go into training and tuning a network for each well is better spent on what a pretrained forecaster does not yet provide: forecasts of pumping operations and other managed stresses, estimates of aquifer properties, and scenarios beyond the range of the observed record. This answers, for one task, the question of what hydrological knowledge contributes once data-driven models predict well (Nearing et al., 2021): it moves from fitting each record to the parts of the problem that records alone cannot resolve. It is a statement about the tested task, levels at monitoring wells over 7 to 90 days with a rainfall forecast, and not about every use of recurrent networks in hydrology.

5.2 Why the adaptation helps

Chronos-2 alone reproduced the recession of the level well but responded weakly and late to rain: an added 100 mm storm raised its forecast by 0.05 m, peaking after 5 days, consistent with the small gain Chronos-2 drew from forecast-day weather in our ten-day study (Kim et al., 2026). After adaptation the response was 0.35 m, peaking after 2 days, the same order and timing as the rises observed in these wells. The adapter therefore mainly teaches the model how Korean aquifers turn rainfall into a change of level, which is why its gain is largest when rainfall for the forecast days is supplied and smaller from past

rainfall alone. Most of the skill gain was reached with a few tens of training wells (Figure 2b), which suggests that this response is largely shared across the stations of one network; a nationwide version would then add value mostly through aquifers and climates not covered here, a hypothesis to be tested with the nationwide data.

5.3 Scope

The evidence comes from one region of the network (29 of the 30 test stations are in the Yeongnam region and 1 in Gangwon-do), where most recharge falls in the June–September monsoon; accuracy in other regions and climates is untested. Every test well had at least a year of training record, so forecasts for wells with shorter records were not tested. The main results assume a perfect rainfall forecast, and operational forecasts or seasonal outlooks will lower the skill of every forecaster that uses them. Pumping schedules and other operations are not inputs, and their effect stays in the error at strongly pumped wells. The forecast intervals of the adapted model are somewhat narrow at long horizons. The model forecasts levels; it does not estimate aquifer properties or replace a groundwater flow model.

6. Conclusion

A time-series foundation model adapted once to the Korean monitoring network forecast the levels of wells at stations it had never seen more accurately than recurrent networks trained on each well’s own record at most wells and at every horizon, with the same input variables and without any training at the new well. Without adaptation the pretrained model showed no significant difference from the per-well LSTM at 30 days. GeoAI-GWFM-KR, its code, a forecasting agent and the benchmark are released for noncommercial use so that agencies and researchers can forecast new wells directly and test the claim on their own records.

Data and code availability

Groundwater levels are available from the National Groundwater Information Center (<https://www.gims.go.kr>) and weather from the Korea Meteorological Administration (<https://data.kma.go.kr>); they are not redistributed. The adapter weights, the benchmark split and the results (CC BY-NC 4.0), and the code and scripts that rebuild every number in this note (PolyForm Noncommercial 1.0.0), are archived at Zenodo (Park et al., 2026a; <https://doi.org/10.5281/zenodo.23584923>) and available at <https://geoaialign.ai/models/gwfm-kr/> and <https://github.com/geoaialignment/geoai-gwfm>. Chronos-2 is available from Amazon under the Apache-2.0 license. Daily rainfall was taken from the weather record paired with each groundwater station in the source data set; the list of paired weather stations is not part of this release.

Competing interests and use of AI tools

The authors are affiliated with GeoAI Alignment, Inc., which develops and releases GeoAI-GWFM-KR. An AI assistant (Claude, Anthropic) was used under the authors’ direction to write and run analysis code and to draft text; the authors designed the study and checked the code, results and text.

Appendix A. Settings and full results

Table A1. Training settings.

Item	GeoAI-GWFM-KR adapter	Per-well networks	Regional LSTM
Base or architecture	Chronos-2 (120 M parameters), LoRA rank 8, scaling 16, on attention q, k, v, o and the output layer	single-layer LSTM or GRU, dense head (hidden $\rightarrow$ hidden, ReLU, dropout 0.1 $\rightarrow$ 1)	single-layer LSTM, same head
Inputs per day	level (gaps as missing), rainfall	level (last observation carried forward), daily change, observed flag, rainfall, 30-, 90-, 365-day rainfall sums	same as per-well
History	up to 1,095 days	120 or 365 days (60 for short records)	120 or 365 days
Training	1,000 steps, batch 64 series, AdamW, learning rate 1×10^{-4} , 90-day windows	Adam 10^{-3} , weight decay 10^{-5} , batch 64, rollout = horizon, learning-rate halving on plateau (8 epochs), early stopping (20 epochs, max 200)	Adam 10^{-3} or 10^{-4} , 40,000 samples per epoch, batch 256, early stopping (8 epochs, max 80)
Loss	Chronos-2 quantile loss	squared error of the cumulative change, scaled by its training standard deviation at each lead, observed days only; plus the fluctuation-equation residual for the physics variant	squared error of the level, scaled by each well’s daily-change standard deviation and the square root of the lead, observed days only

Table A2. Median skill, mean RMSE and the number of test wells (of 46) where each forecaster beats persistence, 2023–2024. “Future rain”: rainfall of the forecast days given; “past rain only”: not given.

Forecaster	Rainfall	7 days	30 days	90 days
GeoAI-GWFM-KR	future rain	0.41 (0.149 m; 46)	0.48 (0.234 m; 46)	0.52 (0.318 m; 46)
GeoAI-GWFM-KR	past rain only	0.07 (0.209 m; 42)	0.16 (0.337 m; 45)	0.19 (0.436 m; 46)
Chronos-2	future rain	0.21 (0.180 m; 45)	0.26 (0.298 m; 46)	0.31 (0.410 m; 46)
Chronos-2	past rain only	0.05 (0.208 m; 44)	0.11 (0.346 m; 44)	0.18 (0.457 m; 44)
Per-well LSTM	future rain	0.37 (0.160 m; 43)	0.30 (0.306 m; 41)	0.33 (0.438 m; 43)
Per-well GRU	future rain	0.41 (0.162 m; 43)	0.33 (0.290 m; 44)	0.37 (0.432 m; 41)
Per-well LSTM + physics loss	future rain	0.39 (0.160 m; 42)	0.29 (0.311 m; 39)	0.34 (0.443 m; 42)
Regional LSTM	future rain	0.25 (0.198 m; 38)	0.20 (0.351 m; 37)	0.24 (0.541 m; 29)

Cells: median skill (mean RMSE; wells beating persistence).

Table A3. Paired comparisons on the 46 test wells (rainfall of the forecast days given unless noted). “First better”: wells where the first forecaster has the lower RMSE. Station-level tests use station means of the paired RMSE differences (30 stations).

Comparison (first vs second)	Horizon	First better	Median skill difference	p	Stations first better	Station-level p
GeoAI- GWFM-KR vs Chronos-2	7 d	44/46	+0.148	< 0.001	28/30	< 0.001
GeoAI- GWFM-KR vs Chronos-2	30 d	45/46	+0.165	< 0.001	29/30	< 0.001
GeoAI- GWFM-KR vs Chronos-2	90 d	46/46	+0.143	< 0.001	30/30	< 0.001
GeoAI- GWFM-KR (past rain only) vs Chronos-2	7 d	23/46	-0.000	0.2	17/30	0.5
GeoAI- GWFM-KR (past rain only) vs Chronos-2	30 d	35/46	+0.031	< 0.001	21/30	0.0066
GeoAI- GWFM-KR (past rain only) vs Chronos-2	90 d	32/46	+0.027	< 0.001	20/30	0.0076
GeoAI- GWFM-KR vs Per-well LSTM	7 d	34/46	+0.029	0.0014	21/30	0.0062
GeoAI- GWFM-KR vs Per-well LSTM	30 d	39/46	+0.128	< 0.001	26/30	< 0.001
GeoAI- GWFM-KR vs Per-well LSTM	90 d	40/46	+0.127	< 0.001	27/30	< 0.001
GeoAI- GWFM-KR vs Per-well GRU	7 d	32/46	+0.016	0.0076	21/30	0.005
GeoAI- GWFM-KR vs Per-well GRU	30 d	38/46	+0.112	< 0.001	25/30	< 0.001

Comparison (first vs second)	Horizon	First better	Median skill difference	p	Stations first better	Station-level p
GeoAI- GWFM-KR vs Per-well GRU	90 d	35/46	+0.097	< 0.001	24/30	< 0.001
GeoAI- GWFM-KR vs Per-well LSTM + physics loss	7 d	31/46	+0.021	0.015	21/30	0.015
GeoAI- GWFM-KR vs Per-well LSTM + physics loss	30 d	42/46	+0.160	< 0.001	28/30	< 0.001
GeoAI- GWFM-KR vs Per-well LSTM + physics loss	90 d	43/46	+0.126	< 0.001	28/30	< 0.001
GeoAI- GWFM-KR vs Regional LSTM	7 d	42/46	+0.167	< 0.001	28/30	< 0.001
GeoAI- GWFM-KR vs Regional LSTM	30 d	46/46	+0.203	< 0.001	30/30	< 0.001
GeoAI- GWFM-KR vs Regional LSTM	90 d	44/46	+0.238	< 0.001	29/30	< 0.001
Chronos-2 vs Per-well LSTM	7 d	7/46	-0.109	< 0.001	5/30	0.0011
Chronos-2 vs Per-well LSTM	30 d	17/46	-0.067	0.084	10/30	0.1
Chronos-2 vs Per-well LSTM	90 d	16/46	-0.039	0.49	10/30	0.43
Chronos-2 vs Regional LSTM	7 d	20/46	-0.028	0.63	14/30	0.53
Chronos-2 vs Regional LSTM	30 d	29/46	+0.052	0.023	20/30	0.035
Chronos-2 vs Regional	90 d	30/46	+0.152	0.002	21/30	0.005

Comparison (first vs second)	Horizon	First better	Median skill difference	p	Stations first better	Station-level p
LSTM						
Regional LSTM vs Per-well LSTM	7 d	9/46	-0.118	< 0.001	6/30	< 0.001
Regional LSTM vs Per-well LSTM	30 d	13/46	-0.096	0.0066	10/30	0.05
Regional LSTM vs Per-well LSTM	90 d	13/46	-0.107	< 0.001	8/30	0.0022
Per-well LSTM + physics loss vs Per-well LSTM (64 units, 120 days)	7 d	20/46	-0.001	0.66	13/30	0.52
Per-well LSTM + physics loss vs Per-well LSTM (64 units, 120 days)	30 d	20/46	-0.003	0.35	12/30	0.49
Per-well LSTM + physics loss vs Per-well LSTM (64 units, 120 days)	90 d	27/46	+0.005	0.43	16/30	0.56

Figures

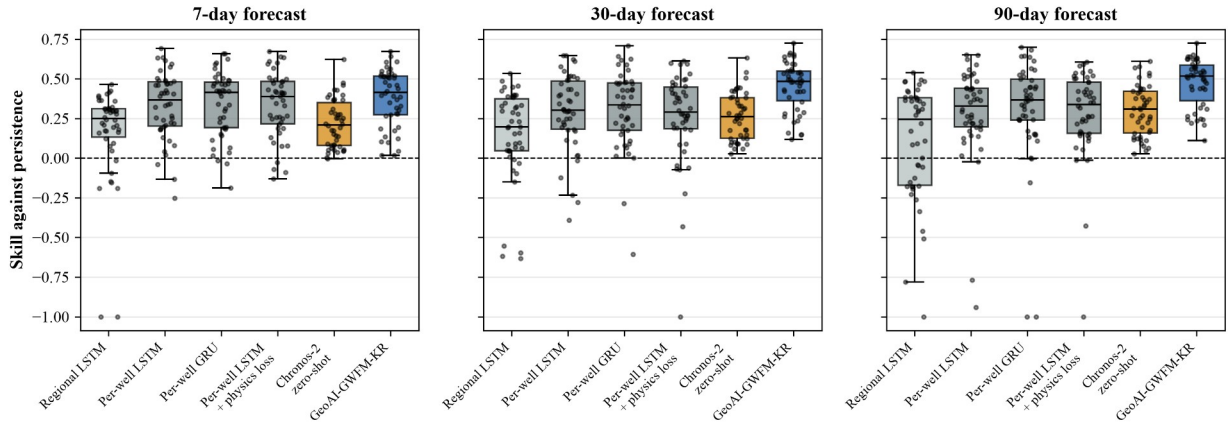


Figure 1. Skill against persistence on the 46 test wells, 2023–2024, by forecaster and horizon, with the rainfall of the forecast days given. One dot per well (values below -1 drawn at -1); boxes span the interquartile range.

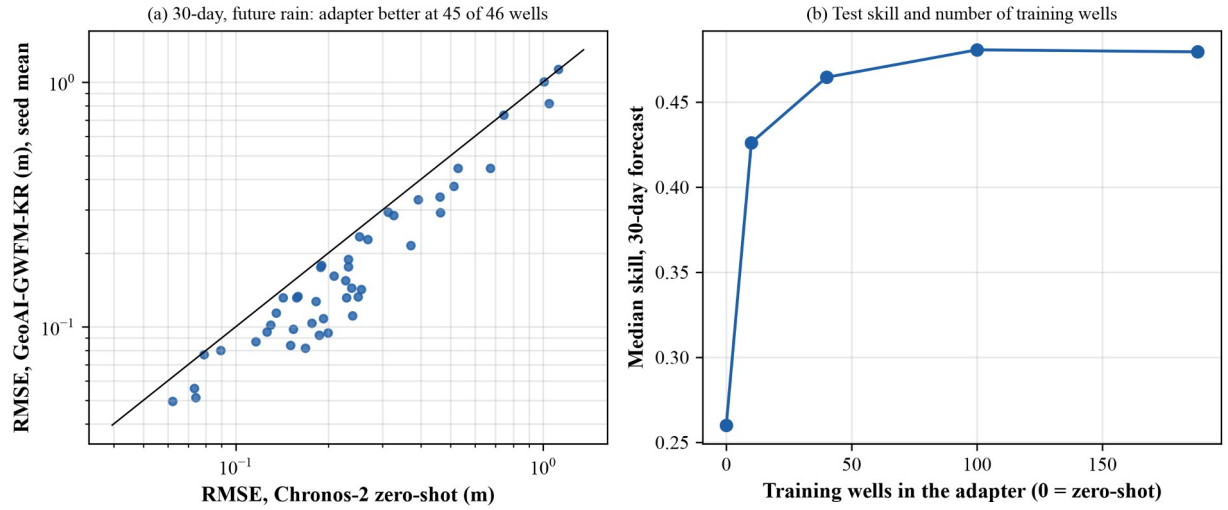


Figure 2. (a) 30-day RMSE at each test well with and without the adapter (adapter averaged over three seeds). (b) Median 30-day skill on the test wells against the number of training wells used for the adapter (seed 0; mean of two random subsets for 10 and 40 wells).

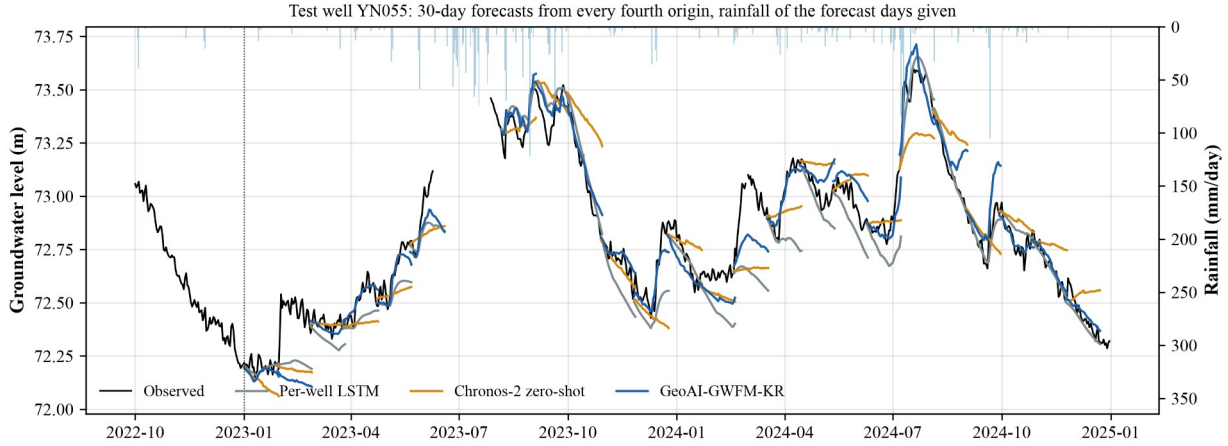


Figure 3. Test well YN055, the well with the median gain of GeoAI-GWFM-KR over the per-well LSTM at 30 days: observed daily level (black), 30-day forecasts from every fourth origin of the per-well LSTM, Chronos-2 alone and GeoAI-GWFM-KR (seed 0), and daily rainfall (bars from the top). The dotted line marks the start of the test period.

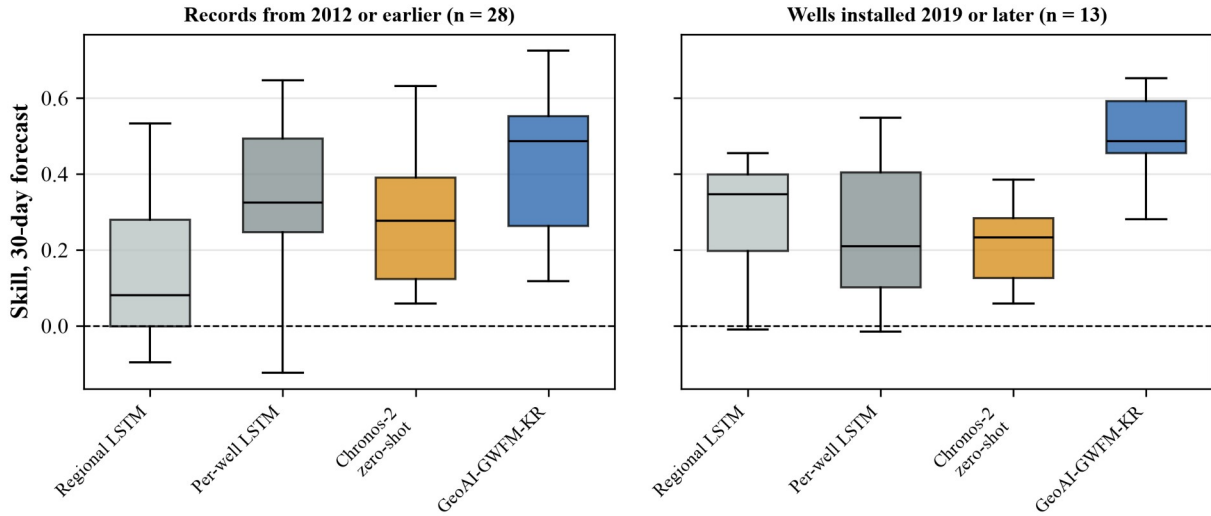


Figure 4. 30-day skill at the 28 test wells with records from 2012 or earlier and the 13 wells installed in 2019 or later.

References

- Aksu, T., Woo, G., Liu, J., Liu, X., Liu, C., et al. (2024). GIFT-Eval: A benchmark for general time series forecasting model evaluation. arXiv:2410.10393. <https://arxiv.org/abs/2410.10393>
- Ansari, A. F., Shchur, O., Küken, J., Auer, A., Han, B., Mercado, P., et al. (2025). Chronos-2: From univariate to universal forecasting. arXiv:2510.15821. <https://arxiv.org/abs/2510.15821>
- Ansari, A. F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., et al. (2024). Chronos: Learning the language of time series. arXiv:2403.07815. <https://arxiv.org/abs/2403.07815>

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1724–1734). <https://doi.org/10.3115/v1/D14-1179>

Das, A., Kong, W., Sen, R., & Zhou, Y. (2024). A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*. arXiv:2310.10688. <https://arxiv.org/abs/2310.10688>

Goswami, M., Szafer, K., Choudhry, A., Cai, Y., Li, S., & Dubrawski, A. (2024). MOMENT: A family of open time-series foundation models. In *Proceedings of the 41st International Conference on Machine Learning*. arXiv:2402.03885. <https://arxiv.org/abs/2402.03885>

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69, 383–393. <https://doi.org/10.1080/01621459.1974.10482962>

Heudorfer, B., Liesch, T., & Broda, S. (2024). On the challenges of global entity-aware deep learning models for groundwater level prediction. *Hydrology and Earth System Sciences*, 28, 525–543. <https://doi.org/10.5194/hess-28-525-2024>

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9, 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., et al. (2022). LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*. <https://arxiv.org/abs/2106.09685>

Kim, T., Park, E., & Park, J. (2026). Sources of skill in short-horizon groundwater-level forecasting: Known future forcing, state memory and site-specific learning in the zero-shot foundation-model era. *ESS Open Archive*. <https://doi.org/10.22541/essoar.15009801/v1>

Korea Meteorological Administration (2026). Daily surface weather observations [Dataset]. Open MET Data Portal. <https://data.kma.go.kr>

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018). Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks. *Hydrology and Earth System Sciences*, 22, 6005–6022. <https://doi.org/10.5194/hess-22-6005-2018>

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, 23, 5089–5110. <https://doi.org/10.5194/hess-23-5089-2019>

Murphy, A. H. (1988). Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly Weather Review*, 116, 2417–2424. [https://doi.org/10.1175/1520-0493\(1988\)116<2417:SSBOTM>2.0.CO;2](https://doi.org/10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2)

National Groundwater Information Center (2026). National Groundwater Monitoring Network daily groundwater levels [Dataset]. <https://www.gims.go.kr>

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., et al. (2021). What role does hydrological science play in the age of machine learning? *Water Resources Research*, 57, e2020WR028091. <https://doi.org/10.1029/2020WR028091>

Park, E., & Parker, J. C. (2008). A simple model for water table fluctuations in response to precipitation. *Journal of Hydrology*, 356, 344–349. <https://doi.org/10.1016/j.jhydrol.2008.04.022>

Park, E., Park, J., & Kim, T. (2026a). GeoAI-GWFM-KR v1.0: a Chronos-2 adapter for groundwater-level forecasting at Korean monitoring wells, with code, benchmark and technical notes (Version 1.0.0) [Software]. Zenodo. <https://doi.org/10.5281/zenodo.23584923>

Park, E., Park, J., & Kim, T. (2026b). What physics still adds to a groundwater-level forecast: record length, recharge timing and pretraining: reproducibility materials (Version v1.0.0-submission) [Software]. Zenodo. <https://doi.org/10.5281/zenodo.23184760>

Read, J. S., Jia, X., Willard, J., Appling, A. P., Zwart, J. A., Oliver, S. K., et al. (2019). Process-guided deep learning predictions of lake water temperature. *Water Resources Research*, 55, 9173–9190. <https://doi.org/10.1029/2019WR024922>

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1, 80–83. <https://doi.org/10.2307/3001968>

Woo, G., Liu, C., Kumar, A., Xiong, C., Savarese, S., & Sahoo, D. (2024). Unified training of universal time series forecasting transformers. arXiv:2402.02592. <https://arxiv.org/abs/2402.02592>

Wunsch, A., Liesch, T., & Broda, S. (2021). Groundwater level forecasting with artificial neural networks: a comparison of long short-term memory (LSTM), convolutional neural networks (CNNs), and non-linear autoregressive networks with exogenous input (NARX). *Hydrology and Earth System Sciences*, 25, 1671–1687. <https://doi.org/10.5194/hess-25-1671-2021>

Wunsch, A., Liesch, T., & Broda, S. (2022). Deep learning shows declining groundwater levels in Germany until 2100 due to climate change. *Nature Communications*, 13, 1221. <https://doi.org/10.1038/s41467-022-28770-2>