

관정별 학습 없이 처음 보는 관정의 지하수위를 예측하는 시계열 기반 모델: 국가지하수관측망에 적응시킨 GeoAI- GWFM-KR

기술 노트, GeoAI-GWFM-KR v1.0

박은규, 박장원, 김태유 (주식회사 지오에이아이얼라인먼트, 대구)

2026년 10월

핵심 요약

- 국내 관측정 188개로 한 번 적응시킨 시계열 기반 모델은 처음 보는 관측소의 관정을 관정별 학습 없이 예측한다.
- 같은 입력 변수로 30일 예측은 46개 관정 중 39개 관정에서 관정마다 학습한 장단기 메모리 (long short-term memory, LSTM) 신경망보다 정확하였다.
- 적응하지 않은 사전학습 모델도 30일과 90일 예측에서 관정별 LSTM과 비교하여 유의한 차이를 보이지 않았다.

요약

관측정의 지하수위 예측 모형은 지금도 대부분 관정마다 그 관정의 기록으로 순환 신경망을 학습하여 만들기 때문에, 새 관정마다 모형을 다시 만들어야 하고 기록이 짧은 관정에서는 성능이 낮다. 사전학습된 시계열 기반 모델(time-series foundation model)은 처음 보는 시계열도 학습 없이 예측하므로, 관정마다 신경망을 학습하는 일이 여전히 정확도를 높이는지 따져 볼 필요가 있다. 본 연구에서는 국가지하수관측망 관측소 중 모형 학습 전에 따로 떼어 둔 30곳 (46개 관정)에서 2023–2024년에 주 단위로 예측을 수행하였으며, 모든 모형에 같은 입력 변수 (관측 수위와 일 강수량)를 주었다. 학습하지 않은 Chronos-2는 관정별로 학습한 장단기 메모리 (long short-term memory, LSTM) 신경망과 비교하여 30일과 90일 예측 모두에서 유의한 차이를 보이지 않았다. 7일 예측에서는 덜 정확하였다. 다른 관측소의 188개 관정으로 저랭크 어댑터를 한 번 학습하여 적응시킨 모형(GeoAI-GWFM-KR)의 persistence 대비 예측력(skill) 중앙값은 7, 30, 90일 예측에서 각각 0.41, 0.48, 0.52으로 관정별 LSTM의 0.37, 0.30, 0.33보다 높았으며, 46개 관정 중 34, 39, 40개 관정에서 더 정확하였다. 2019년 이후 설치된 관정에서는 30일 예측에서 13개 중 13개 관정에서 더 정확하였다. 물리 손실을 더한 관정별 신경망은 정확도가 높아지지 않았다. 강수 예보가 주어지는 7–90일 예측에서, 기록이 1년 이상인 관정에 대해 적응시킨 사전학습 모델 하나는 새 관정에서 학습하지 않고도 관정별 학습과 같거나 더 나은 정확도를 보였다.

1. 서론

오늘 관측정을 새로 설치한 기관은 그 관정의 기록이 쌓이기 훨씬 전부터 앞으로 몇 주의 수위 예측을 필요로 한다. 지금까지는 몇 년간 자료를 모은 뒤 그 관정의 기록으로 장단기 메모리(long short-term memory, LSTM)나 게이트 순환 유닛(gated recurrent unit, GRU) 같은 순환 신경망을 학습하는 방식이 일반적이었으며, 강수를 입력에 넣거나 손실 함수에 물리 관계식을 더하기도 하였다. 이 과정은 관정마다 되풀이되므로 비용이 크고, 새 관정은 여러 해 동안 예측 모형 없이 남으며, 그 결과가 persistence 같은 단순한 기준이나 그 관정을 본 적 없는 모형과 비교되는 경우도 드물다.

자료 기반 지하수위 예측은 대체로 이 방식을 따라 왔다. 순환 신경망(Hochreiter & Schmidhuber, 1997; Cho et al., 2014)은 관정별로 학습한 합성곱 신경망 및 자기회귀 신경망과 비교되었고(Wunsch et al., 2021), 관정마다 하나씩 학습한 합성곱 신경망으로 100개가 넘는 관정의 미래 수위가 전망되기도 하였다(Wunsch et al., 2022). 지표수 수문학에서는 LSTM이 강우-유출 모형으로 높은 성능을 보였으며(Kratzert et al., 2018), 유역 특성과 함께 수백 개 유역으로 학습한 하나의 LSTM이 유역별로 보정한 모형과 대등하거나 그보다 나은 성능을 보였다(Kratzert et al., 2019). 그러나 지하수에서는 이러한 통합 모형을 만들기가 더 어려웠는데, 관정을 구별하기 위한 정적 특성이 여러 관정으로 학습한 모형을 새 관정으로 일반화하지 못하였기 때문이다(Heudorfer et al., 2024). 한편 자료가 부족한 신경망을 돕기 위해 손실 함수에 물리 관계식을 더하는 연구도 이어졌으며(Read et al., 2019), 자료 기반 모형이 잘 예측하게 된 뒤 수문학적 지식이 어떤 역할을 할 것인지는 여전히 열린 질문이다(Nearing et al., 2021).

시계열 기반 모델은 이 출발점을 바꾼다. TimesFM(Das et al., 2024), Moirai(Woo et al., 2024), MOMENT(Goswami et al., 2024), Chronos(Ansari et al., 2024, 2025)와 같은 모델은 다양한 분야의 대규모 시계열로 사전학습되어 있어, 새로운 시계열도 별도의 학습 없이 최근 기록으로 예측하며, Chronos-2처럼 미래에 알려진 공변량을 함께 쓰는 모델도 있다. 이 모델들은 GIFT-Eval(Aksu et al., 2024)과 같은 광범위한 벤치마크에서 자료별로 학습한 모형과 견줄 만한 성능을 보인다. 20년 기록을 가진 국내 관측정 3곳의 10일 예측에서는 학습하지 않은 TimesFM-3가 10년 기록으로 학습하고 조정한 순환 신경망과 대등하거나 그보다 나았고, 단기 예측력의 대부분은 예측 기간의 기상에서 나왔으며, Chronos-2는 그 기상 정보로부터 얻는 개선이 작았다(Kim et al., 2026). 그러나 그대로 쓰거나 한 번 적응시킨 사전학습 모델이 1-3 개월의 예측기간에서, 많은 관정과 한 번도 보지 않은 관측소에 대해 충분히 조정한 관정별 신경망만큼 정확한지는 확인되지 않았다.

본 연구의 목적은 세 가지 질문에 답하는 데 있다. 첫째, 학습하지 않은 사전학습 모델이 처음 보는 관정에서 관정마다 학습한 순환 신경망만큼 정확한가. 둘째, 같은 관측망의 다른 관정으로 사전학습 모델을 한 번 적응시키면 그 답이 달라지는가. 셋째, 각 방식의 장점은 어디에서 나타나는가. 이를 위해 Chronos-2(Ansari et al., 2025)에 국가지하수관측망 자료로 학습한 저랭크 어댑터를 더한 GeoAI-GWFM-KR을 사용하였으며, 어느 쪽에도 유리하지 않도록 입력 변수, 학습과 검증 기간, 채점일을 모든 모형에 같게 하고 시험 관측소는 모든 학습에서 제외하였다.

2. 자료와 벤치마크

관정과 기간. 국가지하수관측망(국가지하수정보센터, National Groundwater Information Center, 2026) 216개 관측소 340개 관정의 일 지하수위와 인근 기상관측소의 일 강수량(기상청, Korea Meteorological Administration, 2026) 2005–2024년 자료를 사용하였다. 자격을 갖춘 관측소 중 148 곳은 영남권(경상북도, 경상남도, 대구, 부산, 울산)에, 5곳은 강원도와 전라북도에 있다. 대부분의 관측소에는 암반 관정이 있으며, 많은 관측소에 충적 관정이나 천부 관정이 함께 있다. 학습 자료는 2021년 12월 31일까지로 하였고, 모든 모형의 설정은 2022년 자료로 골랐으며, 2023–2024년을 시험 기간으로 하였다. 2023년 이전 유효 수위가 2년 이상이고 시험 기간 유효 수위가 1년 이상인 관정을 대상으로 하였으며, 153개 관측소의 234개 관정이 이 조건을 충족하였다.

시험 관측소. 어떤 모형도 실행하기 전에 관측소의 20%를 시·도별로 층화하여 고정된 난수 시드로 떼어 두었다. 이 30개 관측소의 46개 관정(암반 27, 충적 18, 미분류 1)을 시험 관정으로, 나머지 123개 관측소의 188개 관정을 학습 관정으로 사용하였으며, 시험 관정과 같은 관측소에 있는 학습 관정은 없다. 시험 관측소 중 29곳은 영남권에, 1곳은 강원도에 있다. 시험 관정 중 28개는 2012년 이전부터 기록이 있고, 13개는 2019년 이후에 설치되었다.

입력과 채점. 모든 모형은 예측 전날까지의 관측 수위와 일 강수량이라는 같은 변수를 받으며, 그 형태는 각 구조에 맞춘다(Table 1). Chronos-2는 결측을 그대로 둔 원자료와 최대 1,095일의 과거 기록을 받고, 순환 신경망은 결측일에 마지막 관측값을 이어 쓴 수위와 관측 여부, 강수에서 계산한 누적 강수량을 120일 또는 365일의 창으로 받는다. 입력에서는 극단 이상치(2021년까지 자료의 중앙값에서 사분위범위의 10배 이상 벗어난 값, 전체 관측의 0.41%)만 제거하였으므로, 어떤 입력도 이후 시점의 자료에 의존하지 않는다. 채점에는 전체 기록에 대한 품질 검사를 통과한 관측일만 사용하였다. 10일 이상 같은 값이 이어진 구간과 하루짜리 돌출값(7일 이동 중앙값에서 벗어난 정도가 이동 중앙절대편차의 6배를 넘는 값을 찾는 Hampel 형 필터(Hampel, 1974)로 판정하되, 10 mm 이상 강수 후 3일 이내의 상승은 유지)은 채점에서 제외하였으며(전체 관측의 0.61%), 입력에는 그대로 두었다.

예측 과제. 시험 기간 중 7일마다 예측 시점을 두고, 각 모형이 7, 30, 90일 앞의 일 수위를 예측하였다. 주 설정에서는 예측 기간의 관측 강수를 주어 완전한 강수 예보를 대신하였고, 보조 설정에서는 예측 기간의 강수를 주지 않았다. 정확도는 관정별로 채점일 전체의 평균제곱근오차(root-mean-square error, RMSE)로 계산하였으며, 예측력(skill)은 $1 - \text{RMSE}/\text{persistence}$ 의 RMSE로 정의하였다. persistence는 마지막 관측 수위를 그대로 유지하는 예측이다. 이 예측력은 persistence를 기준으로 한 skill score(Murphy, 1988)를 RMSE로 나타낸 것으로, 일반적인 평균제곱오차 형태인 $1 - (\text{RMSE}/\text{persistence})^2$ 보다 보수적이며, 수위 변동 폭이 작은 관정과 큰 관정을 같은 척도에 놓는다. persistence는 예측기간이 길수록 나빠지므로, 예측력은 같은 예측기간 안에서 모형끼리 비교하는 데에만 사용하였다. 한 예측기간의 점수는 1일부터 H일까지의 모든 예측일을 합쳐 계산하였으며, H일째만 채점해도 모형의 순위는 같았다(4.2절). 모든 모형은 같은 날로 채점하였고, 비교는 관정 단위로 짝지어 Wilcoxon 부호순위 검정(Wilcoxon, 1945)을 하였으며, 같은 관측소의 관정은 서로 독립이 아니므로 관측소 평균 차이로 같은 검정을 반복하였다.

3. 예측 모형

Chronos-2와 어댑터. Chronos-2는 시계열의 과거 기록과 과거 및 미래에 알려진 공변량으로 분위수를 예측하는 1.2억 개 파라미터의 인코더 모형이다(Ansari et al., 2025). 사전학습 자료는 공개 예측 벤치마크 모음과 합성 시계열이며, 본 연구의 지하수 기록은 사전학습에 쓰이지 않았다. 최대 1,095일의 과거 기록과 일 강수량을 입력하고 중앙값을 점 예측으로 사용하였다. GeoAI-GWFM-KR은 주의(attention) 투영과 출력층에 계수 8의 저랭크 적응(low-rank adaptation, LoRA; Hu et al., 2022) 행렬을 더한 것으로, 학습 파라미터는 120만 개로 모형의 1.0%이다. 학습 관정은 예측 기간 강수를 아는 형태와 과거 강수만 아는 형태로 두 번씩 학습에 들어가므로, 어댑터 하나가 두 설정을 모두 처리한다. 2021년까지의 학습 기록에서 무작위 위치의 창을 뽑아 90일 예측 길이와 Chronos-2의 분위수 손실로 학습하였다. 학습률(1×10^{-4} 또는 3×10^{-4})과 학습 단계(1,000 또는 3,000)의 네 가지 후보를 학습 관정의 2022년 자료로 평가하였으며, 미리 정한 규칙에 따라 세 예측기간과 두 강수 설정의 예측력 중앙값 평균이 가장 높은 설정을 선택하였다(학습률 1×10^{-4} , 1,000단계; 검증 점수 0.279, 다른 후보 0.21–0.27, 어댑터 없음 0.182). 선택한 설정을 세 개의 시드로 학습하여 결과를 평균하였고, 공개 어댑터는 시험 전에 지정한 시드 0이다.

관정별 순환 신경망. 각 시험 관정에 대해 그 관정의 기록만으로 단층 순환 신경망을 학습하였으며, 구조는 Park et al.(2026b)의 재귀 설계를 따랐다. 신경망은 과거 창(날에 대해 수위(결측일은 마지막 관측값을 이어 씀), 일 변화량, 관측 여부, 일 강수량과 30, 90, 365일 누적 강수량을 읽은 뒤, 하루씩 진행하며 그날의 강수로 일 변화량을 예측하여 수위에 더하고, 예측한 수위를 다시 입력하는 과정을 예측기간만큼 반복한다. 학습은 예측과 같은 길이로 진행하였고, 손실은 관측된 날만으로 계산하였으며, Adam 최적화와 2022년 자료에 의한 조기 종료를 사용하였다. 관정별 LSTM은 은닉 크기(64 또는 128)와 입력 창(120일 또는 365일)을 관정과 예측기간마다 세 시드의 검증 손실로 골랐다. GRU(은닉 64, 창 120일)와, Park and Parker(2008)의 지하수위 변동식을 손실에 더한 LSTM(은닉 64, 창 120일, 손실 가중치는 0.1, 1, 10 중 관정별 선택)도 함께 비교하였다. 선택된 신경망은 46개 관정 중 42개 이상에서 역량 기준(학습 기간 1일 예측 오차가 persistence보다 작음)을 통과하였다. 관정별 신경망은 시험 관정 자신의 과거 기록을 학습하였으며, 어댑터는 이 기록을 보지 않았다.

지역 LSTM. 여러 관정을 함께 학습하는 효과와 사전학습의 효과를 구분하기 위해, 학습 관정 188개 전체로 하나의 LSTM을 관정별 신경망과 같은 입력(각 관정의 2021년까지 자료로 관정별 표준화)으로 학습하여 시험 관측소에 적용하였다. 이 모형은 어댑터와 정확히 같은 학습 자료를 사용한다. 설정은 학습률 10^{-3} 에서 은닉 크기 64 또는 128과 창 120일 또는 365일, 학습률 10^{-4} 에서 은닉 크기 64 또는 128과 창 120일의 여섯 가지 중 검증 손실로 골랐다. 학습률 10^{-4} 의 두 후보는, 이전 입력 설계에서 실행한 10^{-3} 후보들이 처음 몇 회의 반복 안에 최저 검증 손실에 이르고 시험 기간으로 채점된 뒤에 추가하였으며, 여섯 후보 중 최종 선택에는 2022년 검증 손실만 사용하였다(모든 예측기간에서 은닉 64, 창 120일, 학습률 0.0001 선택).

중립적 비교를 위한 설계. Table 1은 각 예측 모형이 받는 정보를, Table A1은 학습 설정을 정리한 것이다. 모든 모형은 관정에 대해 같은 변수, 즉 관측 수위와 과거 및 예측 기간의

강수만 받으며, 달력 특성이나 강수 이외의 기상 자료, 2021년 이후의 학습 자료는 어느 모형에도 주지 않았다. 입력의 형태와 과거 기록 길이는 각 구조에 따른다. 관정별 신경망은 시험 관정 자신의 기록을 학습한다는 점에서 어댑터보다 유리하다. 학습하는 모든 모형은 2022년 자료로 설정을 고르고 세 개의 시드로 학습하였으며, 모든 모형을 같은 날로 채점하였다. 후보 설정의 수는 모형마다 다르므로, 관정별 LSTM의 조정 방식이 비교 결과를 좌우하지 않는지 확인하기 위해 시험 기간 결과를 보고 관정별 LSTM 설정을 사후에 고른 경우도 비교하였다(4.2절).

Table 1. 예측 모형별 입력 정보, 학습 자료와 설정 선택. 모든 설정은 2022년 검증 자료로 골랐으며, 시험 관측소는 사용하지 않았다.

예측 모형	학습 자료	입력	후보 설정	선택 기준	시드
GeoAI-GWFM-KR	학습 관정 188 개, 2021년까지	결측을 포함한 수위(최대 1,095일), 과거와 예측 기간의 강수	4개(학습률 × 학습 단계)	학습 관정의 2022년 예측력 중앙값	3
Chronos-2	없음	GeoAI-GWFM-KR과 같음	없음	없음	결정적
관정별 LSTM	시험 관정 자신, 2021년까지	결측일에 마지막 관측값을 이어 쓴 수위와 관측 여부(120일 또는 365일), 강수와 30·90·365일 누적 강수	4개(은닉 크기 × 창), 관정·예측기간별	검증 손실 (시드 3개 평균)	3
관정별 GRU	시험 관정 자신, 2021년까지	관정별 LSTM과 같음(120일)	1개	조기 종료만	3
관정별 LSTM + 물리 손실	시험 관정 자신, 2021년까지	관정별 LSTM과 같음(120일)	3개(손실 가중치), 관정·예측기간별	검증 손실 (시드 3개 평균)	3
지역 LSTM	학습 관정 188 개, 2021년까지	관정별 LSTM과 같음(120일 또는 365일)	6개(은닉 크기 × 창 × 학습률)	검증 손실 (시드 3개 평균)	3

4. 결과

4.1 학습하지 않은 사전학습 모델

학습하지 않은 Chronos-2의 예측력 중앙값은 7, 30, 90일 예측에서 각각 0.21, 0.26, 0.31이었다 (Table 2, Fig. 1; 평균 오차는 Table A2). 시험 관정 자신의 기록으로 학습한 관정별 LSTM보다 정확한 관정은 46개 중 7, 17, 16개였다($p < 0.001$, $p = 0.084$, $p = 0.49$; 관측소 단위 $p = 0.0011$, $p = 0.1$, $p = 0.43$; 모든 짝 비교는 Table A3). 즉, 관정별 LSTM과 비교하여 30일과 90일 예측 모두에서 유의한 차이를 보이지 않았다. 다만 점추정치는 LSTM 쪽이 더 높았다. 7일

예측에서는 덜 정확하였다. 각 예측기간의 마지막 날만 채점해도 30일과 90일의 차이는 유의하지 않았다($p = 0.84, p = 0.51$). 30일과 90일 예측에서 가장 강한 관정별 신경망인 관정별 GRU와 비교하면, 7일 예측에서는 덜 정확하였고, 30일 예측에서는 덜 정확하였고, 90일 예측에서는 유의한 차이를 보이지 않았다. 어댑터와 같은 188개 관정으로 학습한 지역 LSTM 보다 정확한 관정은 20, 29, 30개였다($p = 0.63, p = 0.023, p = 0.002$).

Table 2. 시험 관정 46개의 persistence 대비 예측력 중앙값(2023–2024년, 예측 기간 강수 제공). 학습하는 모형은 시드 3개의 평균이다.

예측 모형	학습 자료	7일	30일	90일
GeoAI-GWFM-KR	학습 관정 188 개(어댑터)	0.41	0.48	0.52
Chronos-2	없음	0.21	0.26	0.31
관정별 LSTM	시험 관정 자신의 기록	0.37	0.30	0.33
관정별 GRU	시험 관정 자신의 기록	0.41	0.33	0.37
관정별 LSTM + 물리 손실	시험 관정 자신의 기록	0.39	0.29	0.34
지역 LSTM	학습 관정 188 개	0.25	0.20	0.24

4.2 한 번의 적응과 처음 보는 관측소

어댑터를 더하자 예측력 중앙값은 0.41, 0.48, 0.52으로 높아졌으며, GeoAI-GWFM-KR은 46개 관정 중 44, 45, 46개 관정에서 Chronos-2 단독보다 정확하였다(Fig. 2a). 세 시드의 결과는 0.41–0.42, 0.48–0.49, 0.50–0.53이었다. 관정별 LSTM보다 정확한 관정은 34, 39, 40개였다($p = 0.0014, p < 0.001, p < 0.001$; 관측소 단위 $p = 0.0062, p < 0.001, p < 0.001$). 관정별 GRU에 대해서는 32, 38, 35개, 물리 손실을 더한 LSTM에 대해서는 31, 42, 43개, 지역 LSTM에 대해서는 42, 46, 44개 관정에서 더 정확하였다. 두 가지 확인 결과, 이 비교는 채점 방식이나 관정별 LSTM의 조정 방식에 좌우되지 않았다. 각 예측기간의 마지막 날만 채점하면 GeoAI-GWFM-KR은 33, 42, 36개 관정에서 관정별 LSTM보다 정확하였다($p \leq 0.001$). 기준 모형에 가장 유리하도록 시험 기간 결과를 보고 관정별 LSTM 설정을 사후에 고른 경우에도 30일과 90일 예측에서 각각 37, 37개 관정에서 더 정확하였으며($p < 0.001, p < 0.001$), 7일 예측에서는 차이가 유의하지 않았다(46개 중 25개, $p = 0.63$). Fig. 3은 관정별 LSTM 대비 개선이 중앙값인 시험 관정의 예측을 보여 준다.

4.3 관정별 학습이 경쟁력을 보이는 조건

GeoAI-GWFM-KR의 우위는 관정별 신경망이 학습할 수 있는 기록의 길이에 따라 달랐다(Fig. 4). 2019년 이후 설치되어 학습 기록이 1–3년인 13개 관정에서는 7, 30, 90일 예측에서 각각 12, 13, 13개 관정에서 더 정확하였다. 2012년 이전부터 기록이 있는 28개 관정에서는 10년 이상의 기록으로 학습한 관정별 신경망이 강한 예측 모형이었으며, 더 정확한 관정은 28개 중 19, 21, 23개였다($p = 0.13, p < 0.001, p < 0.001$).

4.4 순환 신경망의 설계 선택

순환 신경망의 성능을 높이기 위해 흔히 쓰는 두 가지 선택은 본 연구에서 도움이 되지 않았다. 지하수위 변동식을 손실에 더해도 같은 은닉 크기와 창 의 관정별 LSTM보다 정확도가 달라지지 않았으며(모든 예측기간에서 $p \geq 0.35$), 이는 같은 변동식을 학습 손실로 쓴 선행 연구에서 신경망이 여러 해의 기록을 학습하면 변동식이 더 이상 정보를 더하지 않았던 결과 (Park et al., 2026b)와 같은 방향이다. 학습 관정을 모아 하나의 지역 LSTM으로 학습한 경우에도 관정 자신의 기록을 대신하지 못하여, 지역 LSTM은 46개 관정 중 37, 33, 33개 관정에서 관정별 LSTM보다 덜 정확하였다. 여러 관정의 자료는 사전학습 모델을 적응시키는 데 쓰였을 때에만 도움이 되었다.

4.5 강수에 대한 반응

예측의 강수 반응을 보기 위해 2023–2024년에 28일마다 둔 예측 시점에서 예측 3일째에 100 mm의 강수를 더하였다. GeoAI-GWFM-KR의 30일 예측 중앙값은 0.35 m 상승하여 호우 2일 뒤에 정점에 이르렀고, Chronos-2 단독은 0.05 m 상승하여 5일 뒤에 정점에 이르렀다. 같은 관정들의 2023년 이전 기록에서 하루 80 mm 이상의 강수 뒤 실제 수위 상승의 중앙값은 0.57 m(사분위범위 0.34–0.96 m)였고 정점은 2일 뒤였다. 실제 사건에는 이웃한 날의 강수가 섞여 있어 규모만 비교할 수 있다. 예측 기간 강수를 0, 1, 2배로 바꾸었을 때 30일 뒤 수위가 강수 순서대로 정렬된 비율은 GeoAI-GWFM-KR이 99.5%, Chronos-2 단독이 93.7%였다. 예측 기간의 강수를 주지 않았을 때 어댑터의 예측력 중앙값은 0.07, 0.16, 0.19로 Chronos-2 단독의 0.05, 0.11, 0.18보다 개선 폭이 작았으며, 짝 비교 결과 7일 예측에서는 유의한 차이가 없었고, 30일 예측에서는 유의하게 개선되었고, 90일 예측에서는 유의하게 개선되었다.

4.6 학습 관정 수와 예측 구간

학습 관정 중 무작위로 고른 일부로 어댑터를 학습한 결과(시드 0), 30일 예측력 중앙값은 관정 10개에서 0.43, 40개에서 0.46, 100개에서 0.48, 전체 188개에서 0.48로, 적응하지 않은 경우의 0.26보다 높았다(Fig. 2b). 공개 어댑터의 80% 예측 구간(10–90% 분위수)은 예측 기간 강수를 준 경우 7, 30, 90일 예측에서 관측의 80%, 77%, 74%를, 주지 않은 경우 79%, 72%, 70%를 포함하였으며, Chronos-2 단독은 76–82%를 포함하였다.

5. 토의

5.1 기반 모델 시대에 관정별 학습이 주는 것

관측정마다 순환 신경망을 학습하는 근거는 정확도, 즉 그 관정을 학습한 모형이 그 관정을 가장 잘 예측한다는 데 있었다. 본 벤치마크에서는 기록이 1년 이상인 관정의 7–90일 예측에 대해 관정별 학습이 그 정확도를 주지 못하였다. 이 기록을 한 번도 보지 않은 사전학습 모델은 관정별 LSTM과 비교하여 30일과 90일 예측 모두에서 유의한 차이를 보이지 않았다(점추정치는 LSTM 쪽이 더 높았다). 다른 관측소로 한 번 적응시킨 뒤에는 모든 예측기간에서 대부분의 관정에서 더 정확하였으며, 관정별 학습이 가장 불리한 기록이 짧은 관정에서 차이가 가장 컸다. 관정별 학습은 7일 예측과 10년 이상의 기록을 가진 관정에서 가장 경쟁력이

있었다. 이 관정들의 7일 예측에서 적응시킨 모형은 관정별 LSTM과 비교하여 유의한 차이를 보이지 않았다, 관정별 LSTM 설정을 사후에 고르면 7일 예측의 차이는 사라졌다. 관정별 신경망에 흔히 더하는 물리 손실과 여러 관정의 통합 학습도 이 차이를 줄이지 못하였으며, 통합 학습한 LSTM이 새 관측소로 잘 전이되지 않은 것은 Heudorfer et al.(2024)이 보고한 지하수 통합 모형의 한계와 같은 방향이다.

이 결과의 실무적 의미는 이 과제의 기본 선택지가 바뀐다는 데 있다. 기록이 1년 이상인 신설 관정의 예측은 더 이상 학습 과정을 기다릴 필요가 없으며, 관측망이 공유하는 모형 하나로 예측할 수 있다. 기록이 1년 미만인 관정은 시험하지 않았다. 관정마다 신경망을 학습하고 조정하는 데 쓰던 노력은 사전학습 예측 모형이 아직 제공하지 않는 부분, 즉 양수 운영과 같은 인위적 영향의 예측, 대수층 특성의 추정, 관측 기록 범위를 벗어난 시나리오 분석에 쓰는 것이 바람직하다. 이는 자료 기반 모형이 잘 예측하게 된 뒤 수문학적 지식이 무엇을 기여하는가라는 질문(Nearing et al., 2021)에 대해, 하나의 과제에서 얻은 답이기도 하다. 그 기여는 각 기록을 맞추는 데서, 기록만으로는 풀 수 없는 문제로 옮겨 간다. 이 판단은 시험한 과제, 즉 강수 예보가 주어진 관측점 수위의 7-90일 예측에 대한 것이며, 수문학에서 순환 신경망이 쓰이는 모든 경우에 대한 것은 아니다.

5.2 적응의 효과

Chronos-2 단독은 수위 감수 과정은 잘 재현하였으나 강수에는 약하고 늦게 반응하여, 100 mm 호우를 더해도 예측이 0.05 m 상승하고 5일 뒤에 정점에 이르렀다. 이는 10일 예측 연구에서 Chronos-2가 예측 기간의 기상으로부터 얻은 개선이 작았던 결과(Kim et al., 2026)와 같은 방향이다. 적응 후의 반응은 0.35 m, 2일 뒤 정점으로, 이 관정들에서 관측된 상승과 같은 규모와 시기였다. 따라서 어댑터는 주로 국내 대수층이 강수를 수위 변화로 바꾸는 방식을 모형에 학습시킨 것으로 판단되며, 이 때문에 예측 기간의 강수를 줄 때 이득이 가장 크고 과거 강수만 줄 때는 작다. 예측력 개선의 대부분은 학습 관정 수십 개에서 이미 나타났으므로 (Fig. 2b), 강수 반응은 한 관측망의 관측소들 사이에서 상당 부분 공유되는 것으로 보이며, 전국판은 주로 본 연구가 다루지 않은 대수층과 기후를 포함하는 데에서 가치를 더할 것으로 예상된다. 이 예상은 전국 자료로 검증할 가설이다.

5.3 연구의 범위

본 연구의 근거는 관측망의 한 권역(시험 관측소 30곳 중 29곳이 영남권, 1곳이 강원도), 대부분의 함양이 6-9월 장마철에 일어나는 지역에서 얻은 것이며, 다른 권역과 기후에서의 정확도는 시험하지 않았다. 모든 시험 관정은 1년 이상의 학습 기록을 가졌으므로, 그보다 짧은 기록의 관정은 시험하지 않았다. 주 결과는 완전한 강수 예보를 가정한 것이므로, 실제 예보나 계절 전망을 쓰면 강수를 사용하는 모든 모형의 예측력이 낮아진다. 양수 일정과 같은 운영 정보는 입력에 없으므로, 양수 영향이 큰 관정에서는 그 영향이 예측 오차로 남는다. 적응시킨 모형의 예측 구간은 긴 예측기간에서 다소 좁다. 이 모형은 수위를 예측할 뿐 대수층 특성을 추정하거나 지하수 유동 모형을 대신하지 않는다.

6. 결론

국가지하수관측망에 한 번 적응시킨 시계열 기반 모델은 같은 입력 변수를 받고 새 관정에서 학습하지 않은 상태로, 처음 보는 관측소의 관정 수위를 모든 예측기간에서 대부분의 관정에 대해 관정마다 학습한 순환 신경망보다 정확하게 예측하였다. 적응하지 않은 사전학습 모델도 30일 예측에서 관정별 LSTM과 비교하여 유의한 차이를 보이지 않았다. GeoAI-GWFM-KR과 코드, 예측 에이전트, 벤치마크를 비상업적 이용에 공개하므로, 관련 기관과 연구자는 새 관정을 바로 예측하고 이 결론을 각자의 기록으로 검증할 수 있다.

자료와 코드

지하수위 자료는 국가지하수정보센터(<https://www.gims.go.kr>), 기상 자료는 기상청(<https://data.kma.go.kr>)에서 받을 수 있으며 재배포하지 않는다. 어댑터 가중치, 벤치마크 분할과 결과(CC BY-NC 4.0), 그리고 본 노트의 모든 수치를 재현하는 코드와 스크립트 (PolyForm Noncommercial 1.0.0)는 Zenodo(Park et al., 2026a; <https://doi.org/10.5281/zenodo.23584923>)에 보존하였으며 <https://geoaialign.ai/models/gwfm-kr/> 와 <https://github.com/geoaialignment/geoai-gwfm> 에서도 받을 수 있다. Chronos-2는 Amazon이 Apache-2.0 라이선스로 공개하고 있다. 일 강수량은 원자료에서 각 지하수 관측소에 찍지어진 기상 기록을 사용하였으며, 찍지어진 기상관측소 목록은 이번 공개에 포함하지 않았다.

이해관계와 인공지능 도구 사용

저자들은 GeoAI-GWFM-KR을 개발하여 공개하는 주식회사 지오에이아이얼라인먼트에 소속되어 있다. 분석 코드의 작성과 실행, 원고 초안 작성에는 저자들의 지시에 따라 인공지능 도구(Claude, Anthropic)를 사용하였으며, 연구 설계와 코드, 결과, 원고의 확인은 저자들이 수행하였다.

부록 A. 학습 설정과 전체 결과

Table A1. 학습 설정.

항목	GeoAI-GWFM-KR 어댑터	관정별 신경망	지역 LSTM
기반 또는 구조	Chronos-2(1.2억 파라미터), LoRA 계수 8, 배율 16, 주의 q, k, v, o 투영과 출력층	단층 LSTM 또는 GRU, 완전연결 출력부(은닉 → 은닉, ReLU, dropout 0.1 → 1)	단층 LSTM, 같은 출력부
일별 입력	수위(결측 그대로), 강수	수위(결측은 마지막 관측값), 일 변화량, 관측 여부, 강수, 30·90·365일 누적 강수	관정별과 같음
과거 기록 길이	최대 1,095일	120일 또는 365일(짧은 기록은 60일)	120일 또는 365일
학습	1,000단계, 배치 64개	Adam 10^{-3} , 가중치 감쇠	Adam 10^{-3} 또는 10^{-4} ,

항목	GeoAI-GWFM-KR 어댑터	관정별 신경망	지역 LSTM
	시계열, AdamW, 학습률 1×10^{-4} , 90일 창	10^{-5} , 배치 64, 예측기간과 같은 재귀 길이, 정제 시 학습률 절반(8회), 조기 종료(20회, 최대 200회)	반복당 표본 40,000개, 배치 256, 조기 종료(8회, 최대 80회)
손실	Chronos-2 분위수 손실	누적 변화량의 제공 오차(예측일별 학습 표준편차로 표준화, 관측일만), 물리 변형은 변동식 잔차 추가	수위의 제공 오차(관정별 일 변화량 표준편차와 예측일의 제공근으로 표준화, 관측일만)

Table A2. 예측 모형별 예측력 중앙값, 평균 RMSE, persistence보다 정확한 시험 관정 수(46개 중), 2023–2024년. “예보 강수”는 예측 기간 강수를 준 경우, “과거 강수만”은 주지 않은 경우이다.

예측 모형	강수	7일	30일	90일
GeoAI-GWFM-KR	예보 강수	0.41 (0.149 m; 46)	0.48 (0.234 m; 46)	0.52 (0.318 m; 46)
GeoAI-GWFM-KR	과거 강수만	0.07 (0.209 m; 42)	0.16 (0.337 m; 45)	0.19 (0.436 m; 46)
Chronos-2	예보 강수	0.21 (0.180 m; 45)	0.26 (0.298 m; 46)	0.31 (0.410 m; 46)
Chronos-2	과거 강수만	0.05 (0.208 m; 44)	0.11 (0.346 m; 44)	0.18 (0.457 m; 44)
관정별 LSTM	예보 강수	0.37 (0.160 m; 43)	0.30 (0.306 m; 41)	0.33 (0.438 m; 43)
관정별 GRU	예보 강수	0.41 (0.162 m; 43)	0.33 (0.290 m; 44)	0.37 (0.432 m; 41)
관정별 LSTM + 물리 손실	예보 강수	0.39 (0.160 m; 42)	0.29 (0.311 m; 39)	0.34 (0.443 m; 42)
지역 LSTM	예보 강수	0.25 (0.198 m; 38)	0.20 (0.351 m; 37)	0.24 (0.541 m; 29)

셀: 예측력 중앙값(평균 RMSE; persistence보다 정확한 관정 수).

Table A3. 시험 관정 46개의 짝 비교(별도 표시가 없으면 예측 기간 강수 제공). “앞 모형이 정확”은 앞 모형의 RMSE가 더 작은 관정 수이다. 관측소 단위 검정은 관측소별 RMSE 차이의 평균을 사용하였다(관측소 30 곳).

비교(앞 모형 대 뒤 모형)	예측기간	앞 모형이 정확	예측력 차이 중앙값	p	관측소 단위: 앞 모형이 정확	관측소 단위 p
GeoAI-GWFM-KR 대 Chronos-2	7일	44/46	+0.148	< 0.001	28/30	< 0.001
GeoAI-GWFM-KR 대 Chronos-2	30일	45/46	+0.165	< 0.001	29/30	< 0.001
GeoAI-GWFM-KR 대 Chronos-2	90일	46/46	+0.143	< 0.001	30/30	< 0.001

비교(앞 모형 대 뒤 모형)	예측기간	앞 모형이 정확	예측력 차이 증양값	p	관측소 단위: 앞 모형이 정확	관측소 단위 p
Chronos-2						
GeoAI- GWFM- KR(과거 강수만) 대 Chronos-2	7일	23/46	-0.000	0.2	17/30	0.5
GeoAI- GWFM- KR(과거 강수만) 대 Chronos-2	30일	35/46	+0.031	< 0.001	21/30	0.0066
GeoAI- GWFM- KR(과거 강수만) 대 Chronos-2	90일	32/46	+0.027	< 0.001	20/30	0.0076
GeoAI- GWFM-KR 대 관정별 LSTM	7일	34/46	+0.029	0.0014	21/30	0.0062
GeoAI- GWFM-KR 대 관정별 LSTM	30일	39/46	+0.128	< 0.001	26/30	< 0.001
GeoAI- GWFM-KR 대 관정별 LSTM	90일	40/46	+0.127	< 0.001	27/30	< 0.001
GeoAI- GWFM-KR 대 관정별 GRU	7일	32/46	+0.016	0.0076	21/30	0.005
GeoAI- GWFM-KR 대 관정별 GRU	30일	38/46	+0.112	< 0.001	25/30	< 0.001
GeoAI- GWFM-KR 대 관정별 GRU	90일	35/46	+0.097	< 0.001	24/30	< 0.001
GeoAI- GWFM-KR 대 관정별 LSTM + 물리 손실	7일	31/46	+0.021	0.015	21/30	0.015
GeoAI- GWFM-KR 대 관정별	30일	42/46	+0.160	< 0.001	28/30	< 0.001

비교(앞 모형 대 뒤 모형)	예측기간	앞 모형이 정확	예측력 차이 증양값	p	관측소 단위: 앞 모형이 정확	관측소 단위 p
LSTM + 물리 손실						
GeoAI- GWFM-KR 대 관정별 LSTM + 물리 손실	90일	43/46	+0.126	< 0.001	28/30	< 0.001
GeoAI- GWFM-KR 대 지역 LSTM	7일	42/46	+0.167	< 0.001	28/30	< 0.001
GeoAI- GWFM-KR 대 지역 LSTM	30일	46/46	+0.203	< 0.001	30/30	< 0.001
GeoAI- GWFM-KR 대 지역 LSTM	90일	44/46	+0.238	< 0.001	29/30	< 0.001
Chronos-2 대 관정별 LSTM	7일	7/46	-0.109	< 0.001	5/30	0.0011
Chronos-2 대 관정별 LSTM	30일	17/46	-0.067	0.084	10/30	0.1
Chronos-2 대 관정별 LSTM	90일	16/46	-0.039	0.49	10/30	0.43
Chronos-2 대 지역 LSTM	7일	20/46	-0.028	0.63	14/30	0.53
Chronos-2 대 지역 LSTM	30일	29/46	+0.052	0.023	20/30	0.035
Chronos-2 대 지역 LSTM	90일	30/46	+0.152	0.002	21/30	0.005
지역 LSTM 대 관정별 LSTM	7일	9/46	-0.118	< 0.001	6/30	< 0.001
지역 LSTM 대 관정별 LSTM	30일	13/46	-0.096	0.0066	10/30	0.05
지역 LSTM 대 관정별 LSTM	90일	13/46	-0.107	< 0.001	8/30	0.0022
관정별 LSTM + 물리 손실 대 관정별	7일	20/46	-0.001	0.66	13/30	0.52

비교(앞 모형 대 뒤 모형)	예측기간	앞 모형이 정확	예측력 차이 중앙값	p	관측소 단위: 앞 모형이 정확	관측소 단위 p
LSTM(은닉 64, 창 120 일)						
관정별	30일	20/46	-0.003	0.35	12/30	0.49
LSTM + 물리 손실 대 관정별						
LSTM(은닉 64, 창 120 일)						
관정별	90일	27/46	+0.005	0.43	16/30	0.56
LSTM + 물리 손실 대 관정별						
LSTM(은닉 64, 창 120 일)						

그림

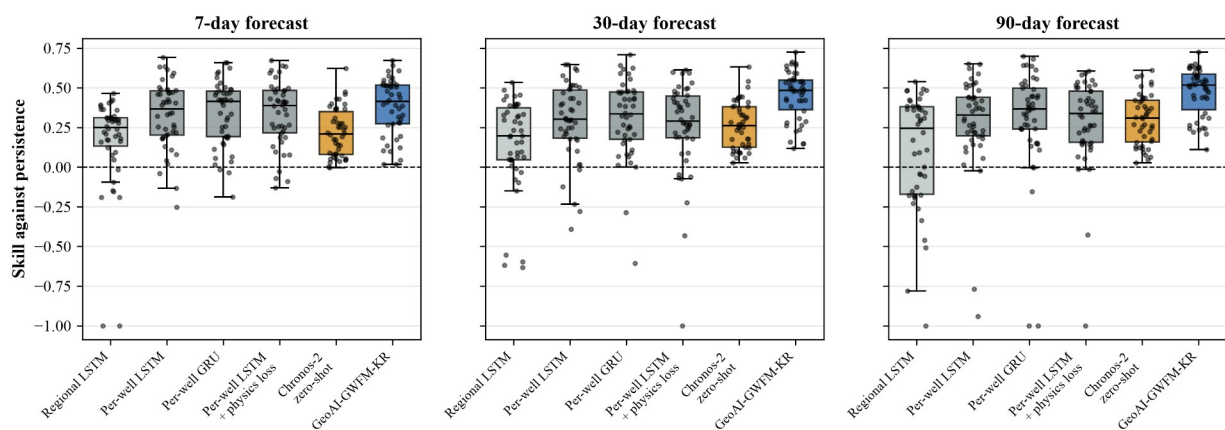


Figure 1. Skill against persistence on the 46 test wells, 2023–2024, by forecaster and horizon, with the rainfall of the forecast days given. One dot per well (values below -1 drawn at -1); boxes span the interquartile range.

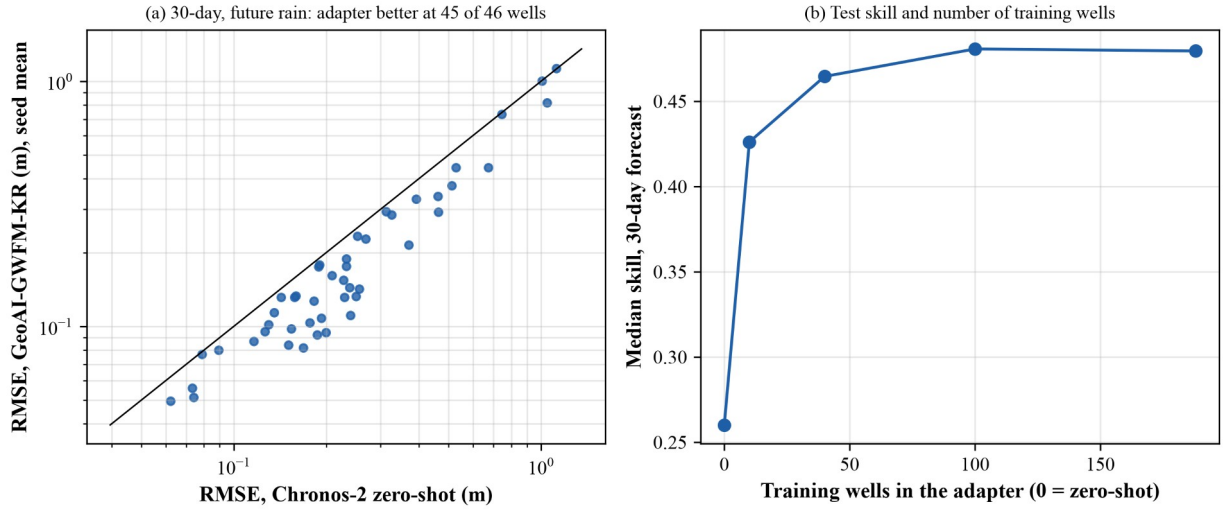


Figure 2. (a) 30-day RMSE at each test well with and without the adapter (adapter averaged over three seeds). (b) Median 30-day skill on the test wells against the number of training wells used for the adapter (seed 0; mean of two random subsets for 10 and 40 wells).

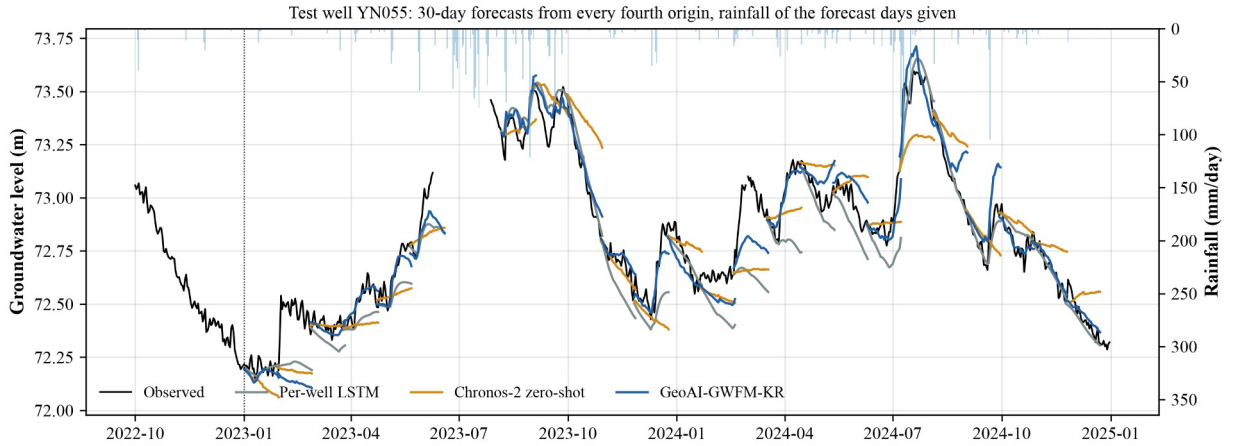


Figure 3. Test well YN055, the well with the median gain of GeoAI-GWFM-KR over the per-well LSTM at 30 days: observed daily level (black), 30-day forecasts from every fourth origin of the per-well LSTM, Chronos-2 alone and GeoAI-GWFM-KR (seed 0), and daily rainfall (bars from the top). The dotted line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

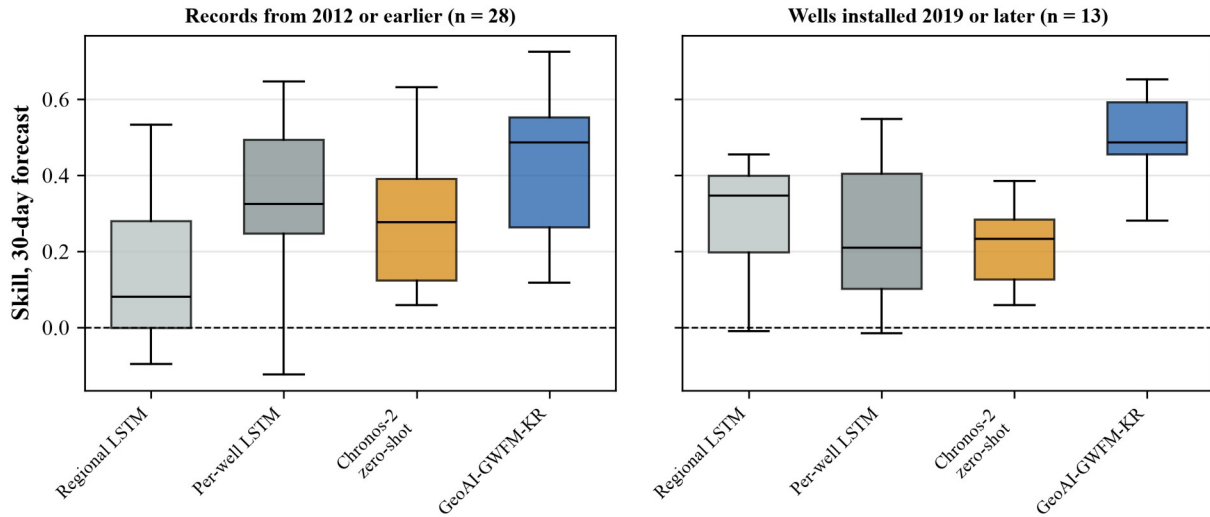


Figure 4. 30-day skill at the 28 test wells with records from 2012 or earlier and the 13 wells installed in 2019 or later.

참고문헌

- 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>