Supplementary Materials for paper 'Exploring the generalisation ability and interpretability of Long Short-Term Memory (LSTM) networks for large-sample groundwater level predictions'

Qidong Fang[1], Mostaquimur Rahman[1], Thorsten Wagener[2], Francesca Pianosi[1]

[1]School of Civil, Aerospace and Design Engineering, University of Bristol, Bristol, BS8 1US, UK
[2]Institute of Environmental Science and Geography, University of Potsdam, Potsdam, 14476, Germany

Correspondence to: Qidong Fang (Qidong.fang@bristol.ac.uk)

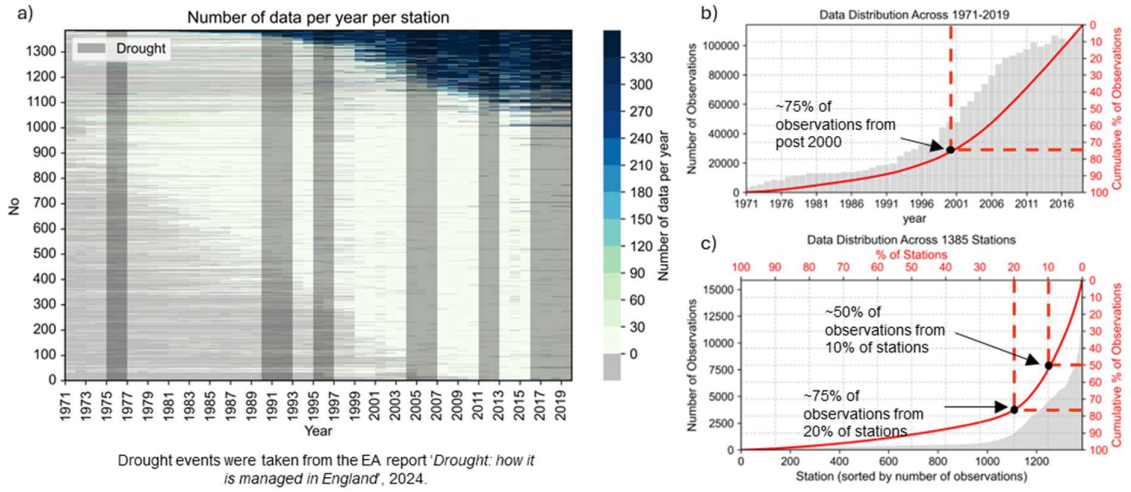


Fig. S1. Overview of the groundwater dataset over time and across stations. (a) Distribution of groundwater observations over time across 1384 stations. Colour indicates the by number of observations per year, while light grey denotes years with no data. Dark grey shading marks the drought events that occurred during the 1975 - 2019, based on records from the Environment Agency (https://www.gov.uk/government/publications/drought-management-for-england/drought-how-it-is-managed-in-england#drought-in-england-an-overview). (b) Distribution of observations by year. (c) Distribution of observations by station.

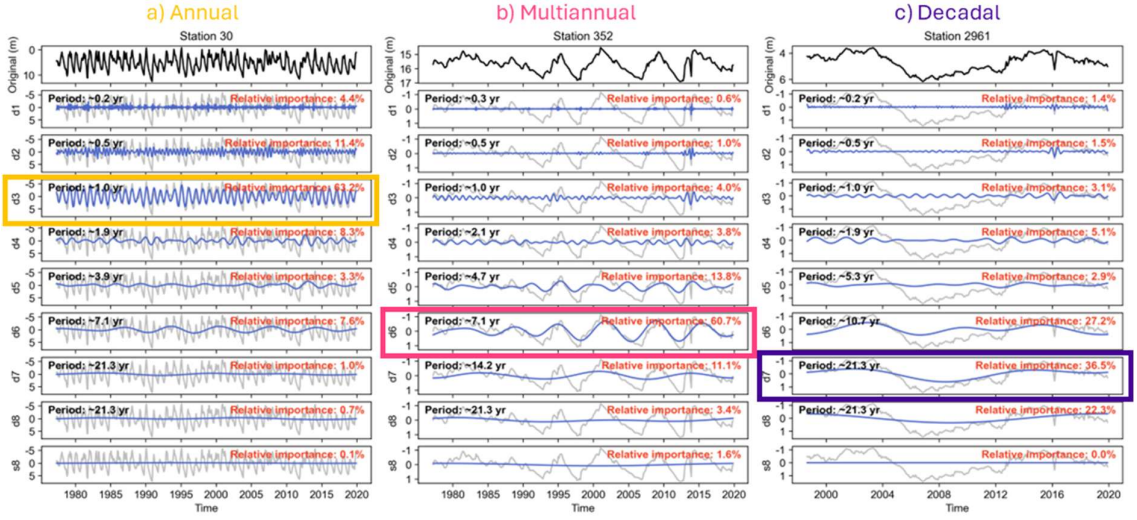


Fig. S2. Examples of wavelet transform decomposition of monthly groundwater level timeseries. (a) Example of a station dominated by an annual cycle. (b) Example of a station dominated by a multi-annual cycle.

(c) Station dominated by a decadal cycle. In each panel, the black line at the top and the grey lines show the original groundwater timeseries. The blue lines represent the wavelet detail components (d1-d8) and the final approximation component (s8). The period (in black) and relative importance (in red) associated with each detail can be calculated. The bold coloured box highlights the dominant fluctuation period for each example.

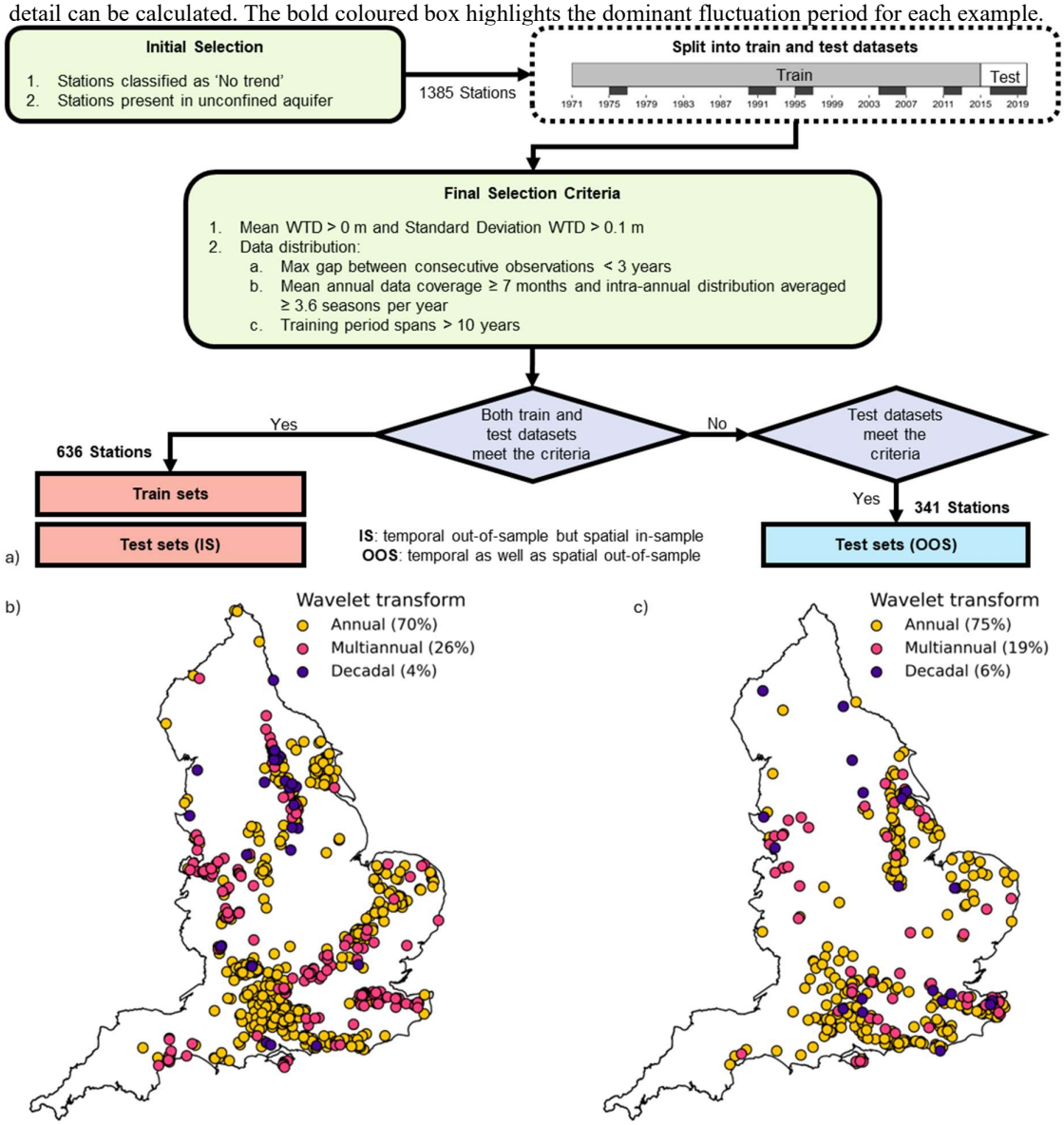


Fig. S3. Schematic of groundwater dataset construction for model training and in-sample / out-of-sample testing, and spatial distribution of stations. (a) Workflow used to construct the groundwater dataset for training, in-sample testing, and out-of-sample testing. (b) Map of groundwater stations included in the train and in-sample test sets. (c) Map of groundwater stations included in the out-of-sample test set. Stations in (b) and (c) are classified based on wavelet transform analysis, indicating whether groundwater level is primarily influenced by annual (yellow), multi-annual (pink) or decadal (purple) cycles.

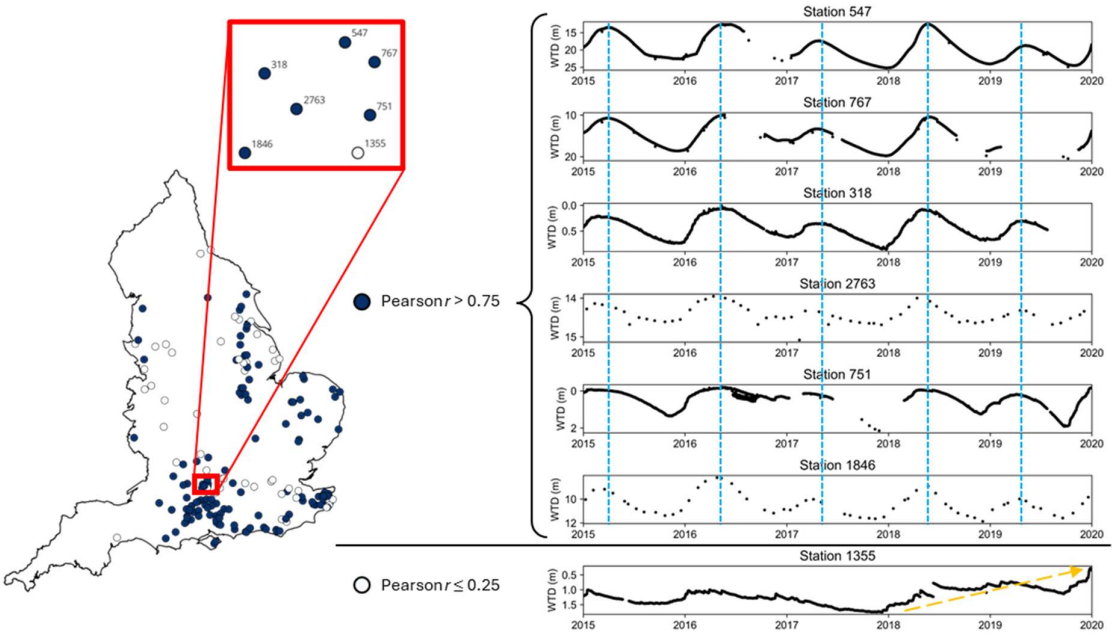


Fig. S4. Representative nearby stations showing opposite performance extremes: Pearson $r \leq 0.25$ (poor performance) and Pearson $r > 0.75$ (good performance, predictable). The map and zoomed-in area (red rectangle) show the locations of all qualified stations and seven example stations. Dark circles represent stations achieving good performance, and white circles represent stations achieving poor performance. In the right panel (top six), we showed six timeseries for stations achieving good performance (stations 547, 767, 318, 2763, 751, and 1846), which exhibit clear seasonal variation (blue dashed lines denote the approximate annual peaks of predictable timeseries). In the right panel bottom, we showed one timeseries for a station achieving poor performance (station 1355), where the station does not exhibit clear seasonal variation and even shows an increase during 2018 – 2020 (yellow arrow).

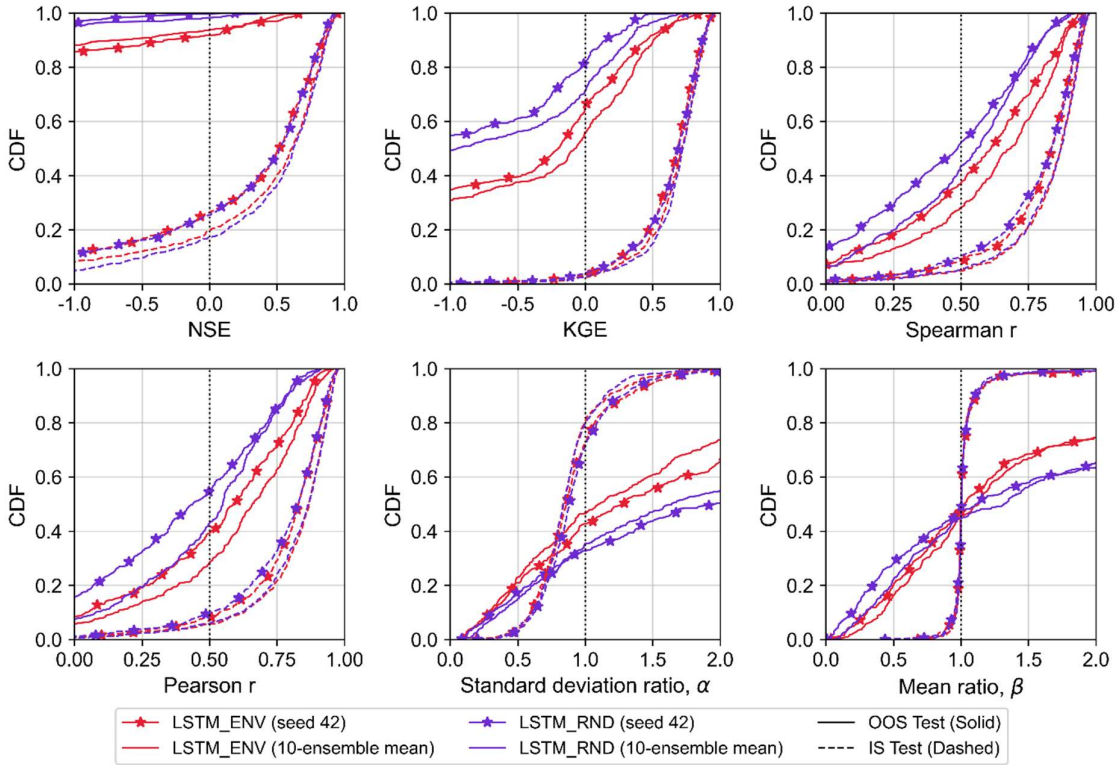


Fig. S5. Cumulative distribution functions (CDFs) of performance metrics for LSTM_ENV (red) and LSTM_RND (purple). Comparisons are shown between the 10-ensemble mean (no markers) and one of the

ensemble member (star markers). Performances is shown for 636 in-sample stations (dashed lines) and 341 out-of-sample stations (i.e., stations not used for the LSTM training, solid lines). For NSE, KGE, Spearman *r*, and Pearson *r*, superior performance is indicated by curves shifted toward the bottom-right. For the Standard deviation ratio (α) and Mean ratio (β), proximity to the vertical (ratio=1) indicates higher accuracy.

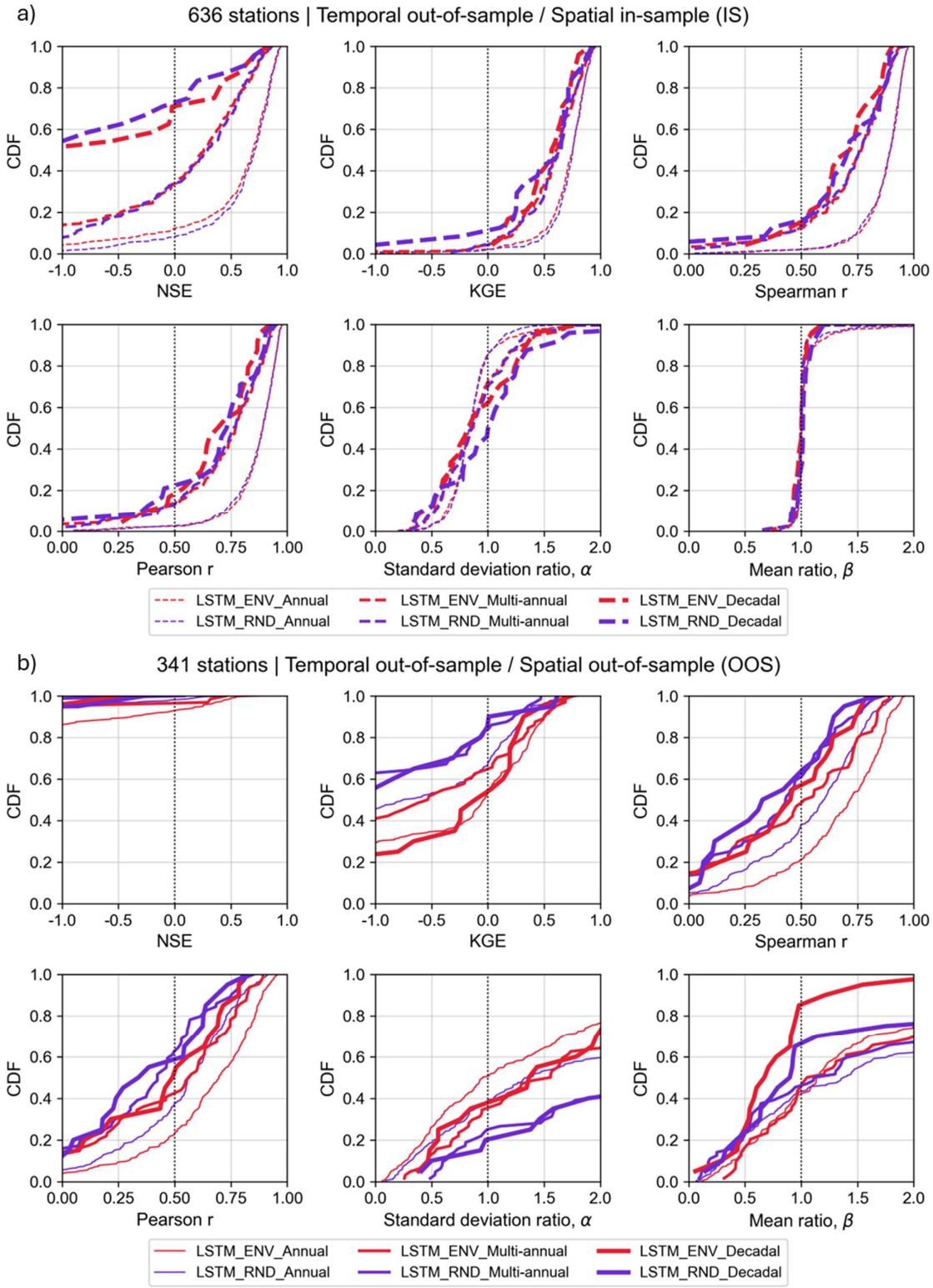


Fig. S6. Cumulative distribution functions (CDFs) of performance metrics for the 10-ensemble mean of LSTM_ENV (red) and LSTM_RND (purple), categorised by dominant variation cycle (i.e. annual, multi-

annual, and decadal) of the stations. Performances is shown for 636 in-sample stations (a) and 341 out-of-sample stations (b). For NSE, KGE, Spearman *r*, and Pearson *r*, superior performance is indicated by curves shifted toward the bottom-right. For the Standard deviation ratio (α) and Mean ratio (β), proximity to the vertical (ratio=1) indicates higher accuracy.

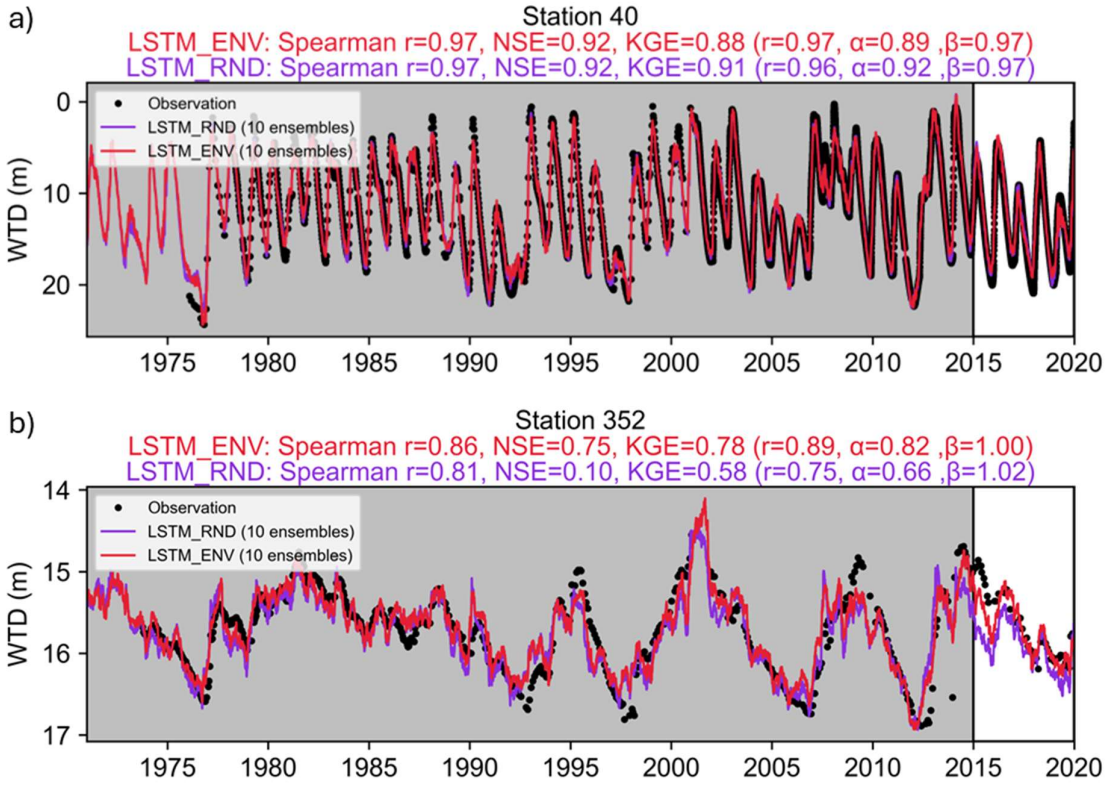


Fig. S7. Representative water table depth (WTD) timeseries for spatially in-sample stations. (a) Station 40, exhibiting a dominant annual cycle, and (b) Station 352, exhibiting a dominant multi-annual cycle. Model simulations and performance metrics for LSTM_ENV and LSTM_RND are shown in red and purple, respectively. Black dots represent observations. The grey background indicates the training period (1971-2014), followed by the tested period (2015-2019).

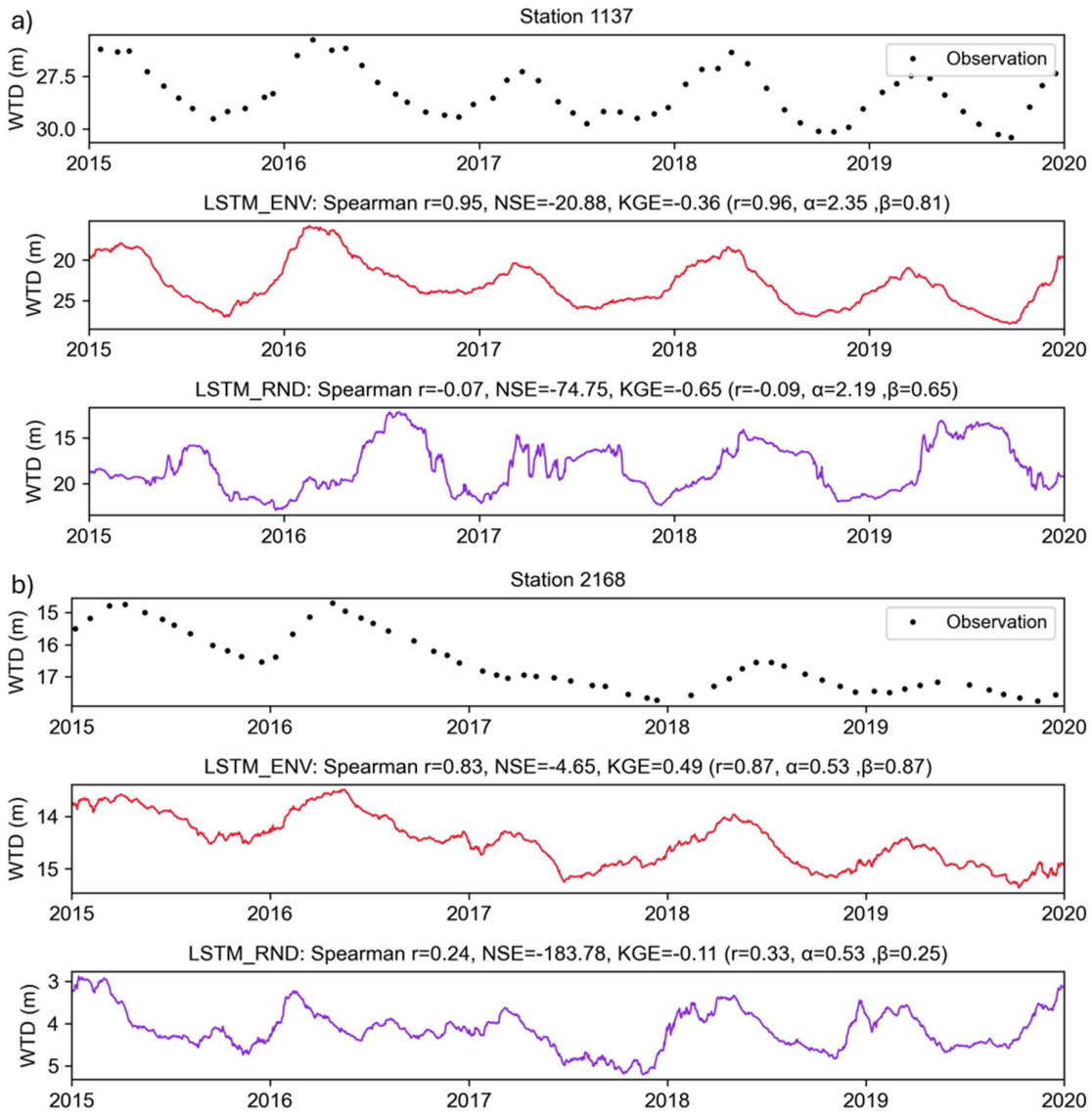


Fig. S8. Representative water table depth (WTD) timeseries for spatially out-of-sample stations. (a) Station 1137, exhibiting a dominant annual cycle, and (b) Station 2168, exhibiting a dominant multi-annual cycle according to wavelet transform analysis. Model simulations and performance metrics for LSTM_ENV and LSTM_RND are shown in red and purple, respectively. Black dots represent observations. Simulations and observations are plotted on separate vertical axes to better compare temporal dynamics, as the means and ranges of observations and simulations differ substantially.

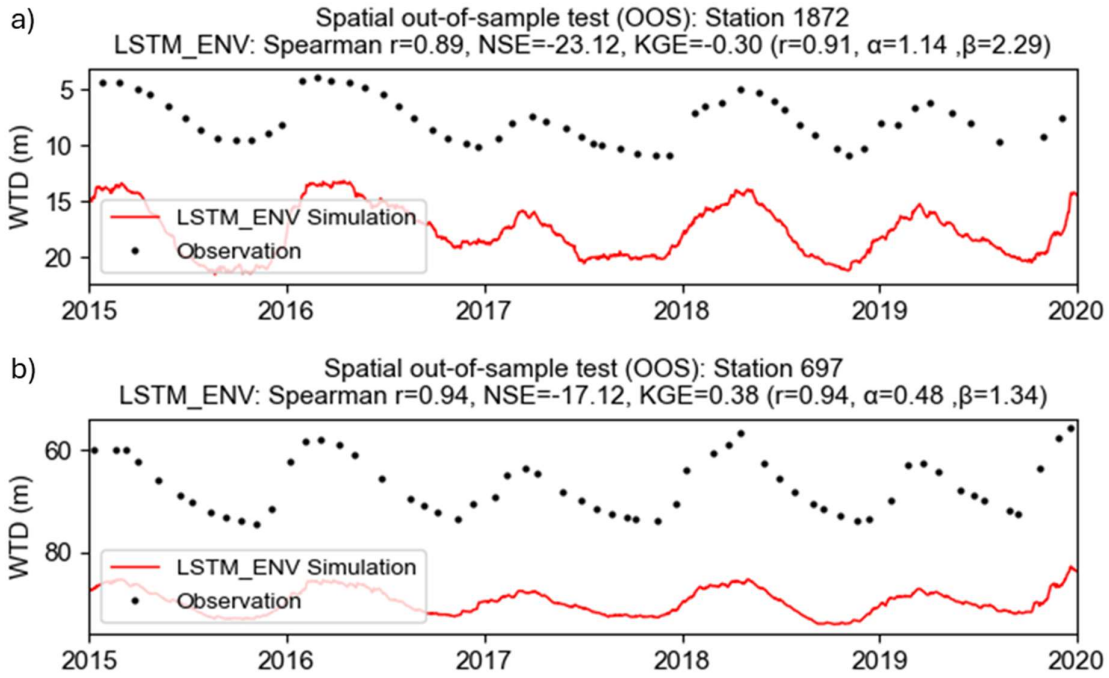


Fig. S9. Representative water table depth (WTD) stations that achieved high correlation but negative NSE. (a) Station 1872, and (b) Station 697.

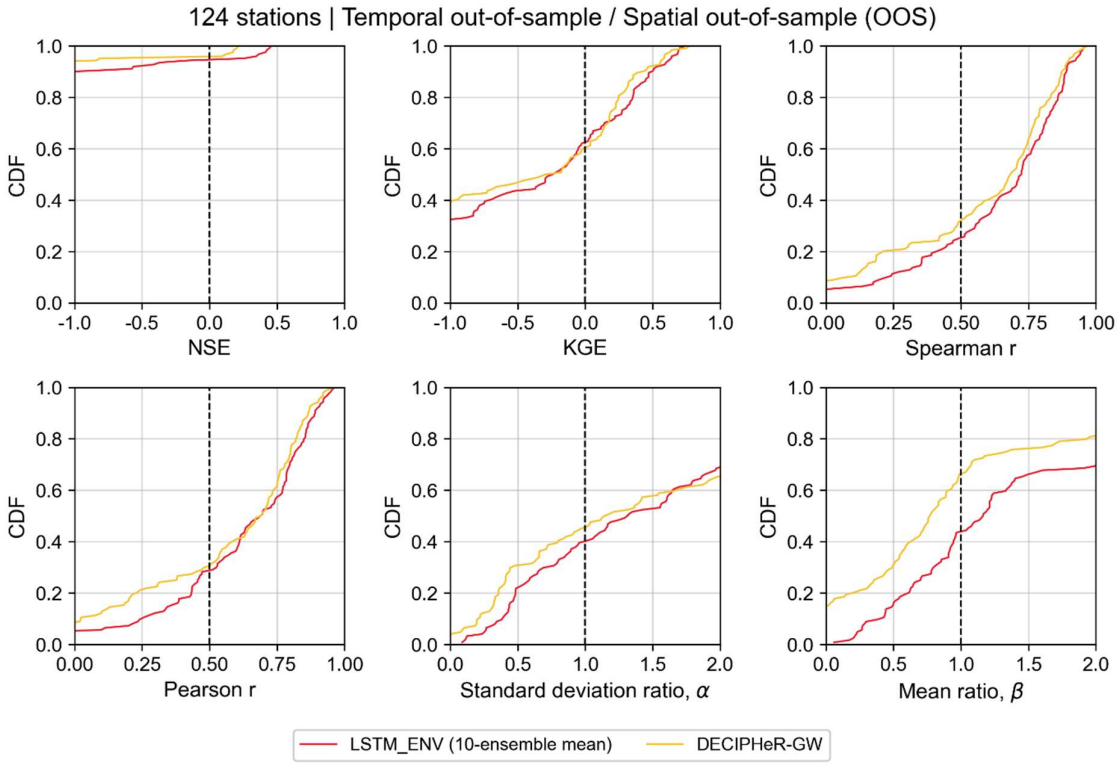


Fig. S10. Cumulative distribution functions (CDFs) of performance metrics for LSTM_ENV (red) and DECIPHeR-GW (yellow). CDFs compile the results from 124 out-of-sample stations. For NSE, KGE,

Spearman *r*, and Pearson *r*, superior performance is indicated by curves shifted toward the bottom-right. For Standard deviation ratio (α) and mean ratio (β), proximity to the vertical (ratio=1) indicates higher accuracy.

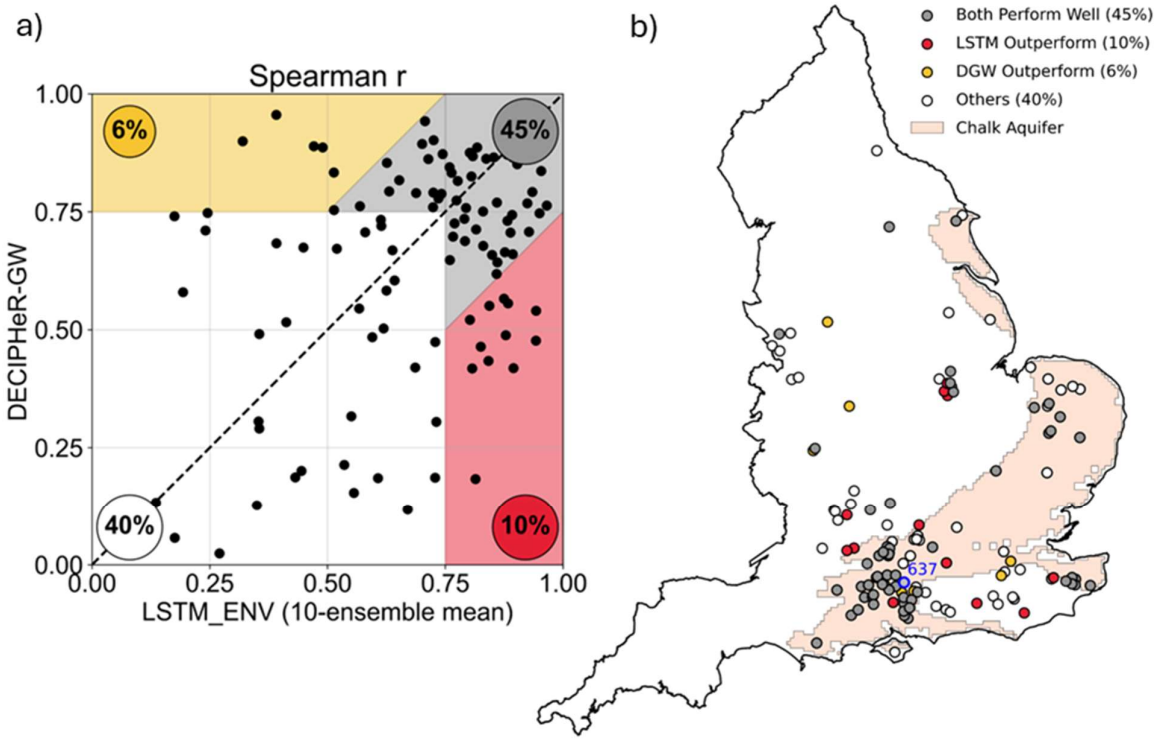


Fig. S11. Performance comparison (Spearman correlation coefficient) between LSTM_ENV (red) and DECIPHeR-GW (yellow) across 124 common stations. (a) Comparison of Spearman correlation coefficient. (b) Spatial distribution of performance classes. Stations (circles) are categorised by their relative Spearman *r*: grey indicates high performance for both models (Spearman *r* of one model > 0.75 and the difference between models $\Delta r < 0.25$); yellow indicates that DECIPHeR-GW outperforms LSTM_ENV, red indicates that LSTM_ENV outperforms DECIPHeR-GW, and white represents all remaining cases. The Chalk aquifer is shown in the background.

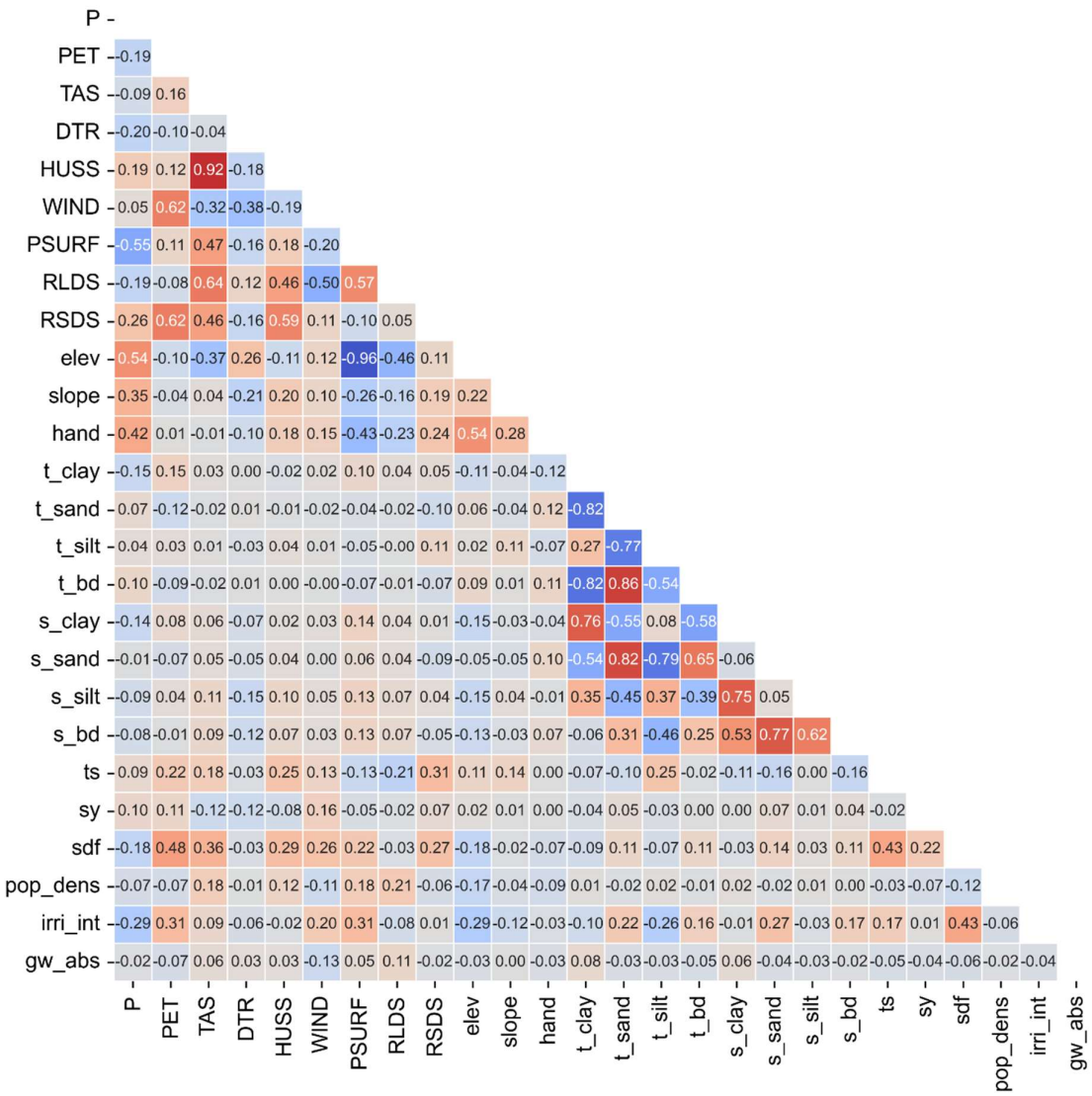
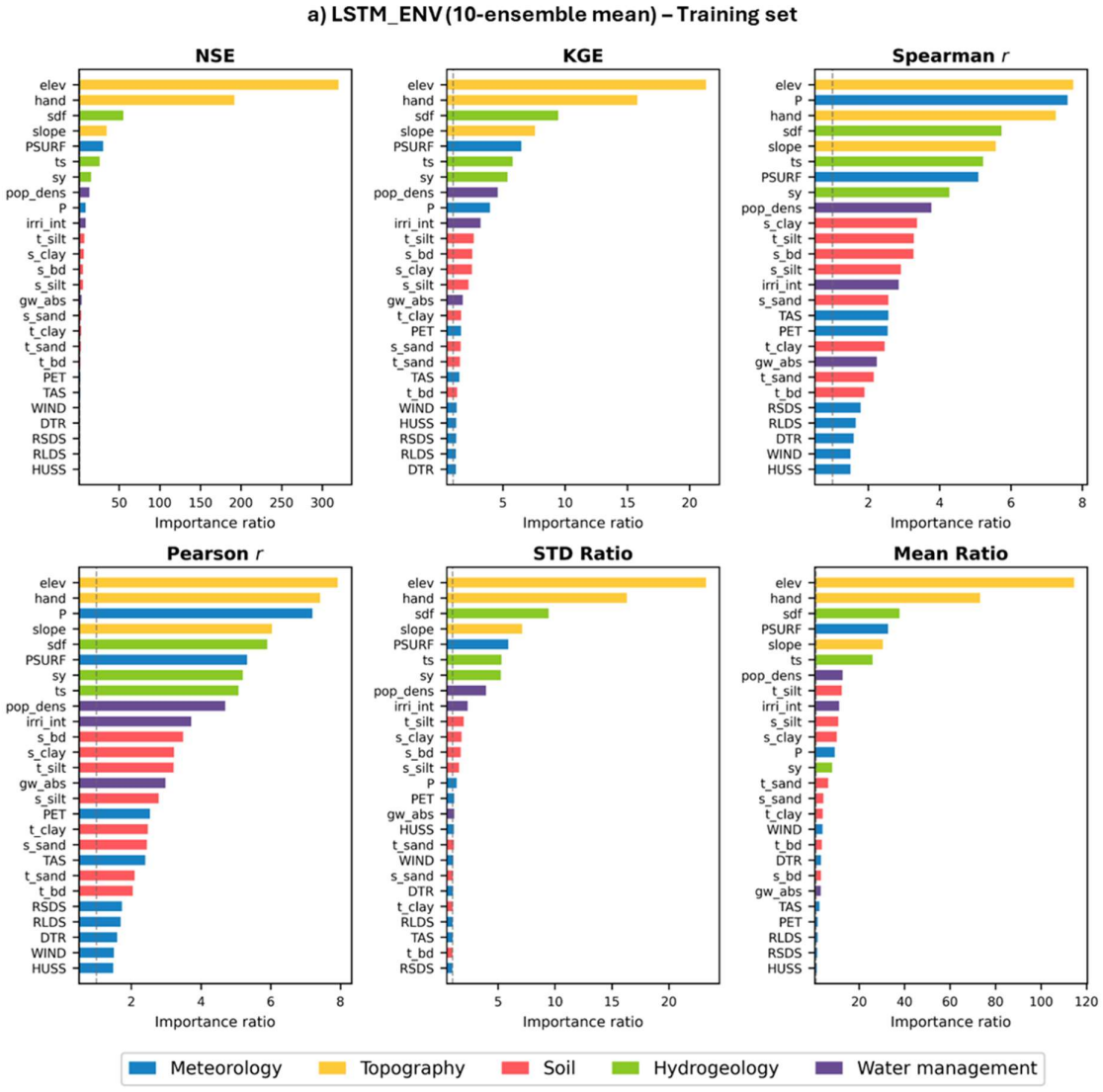
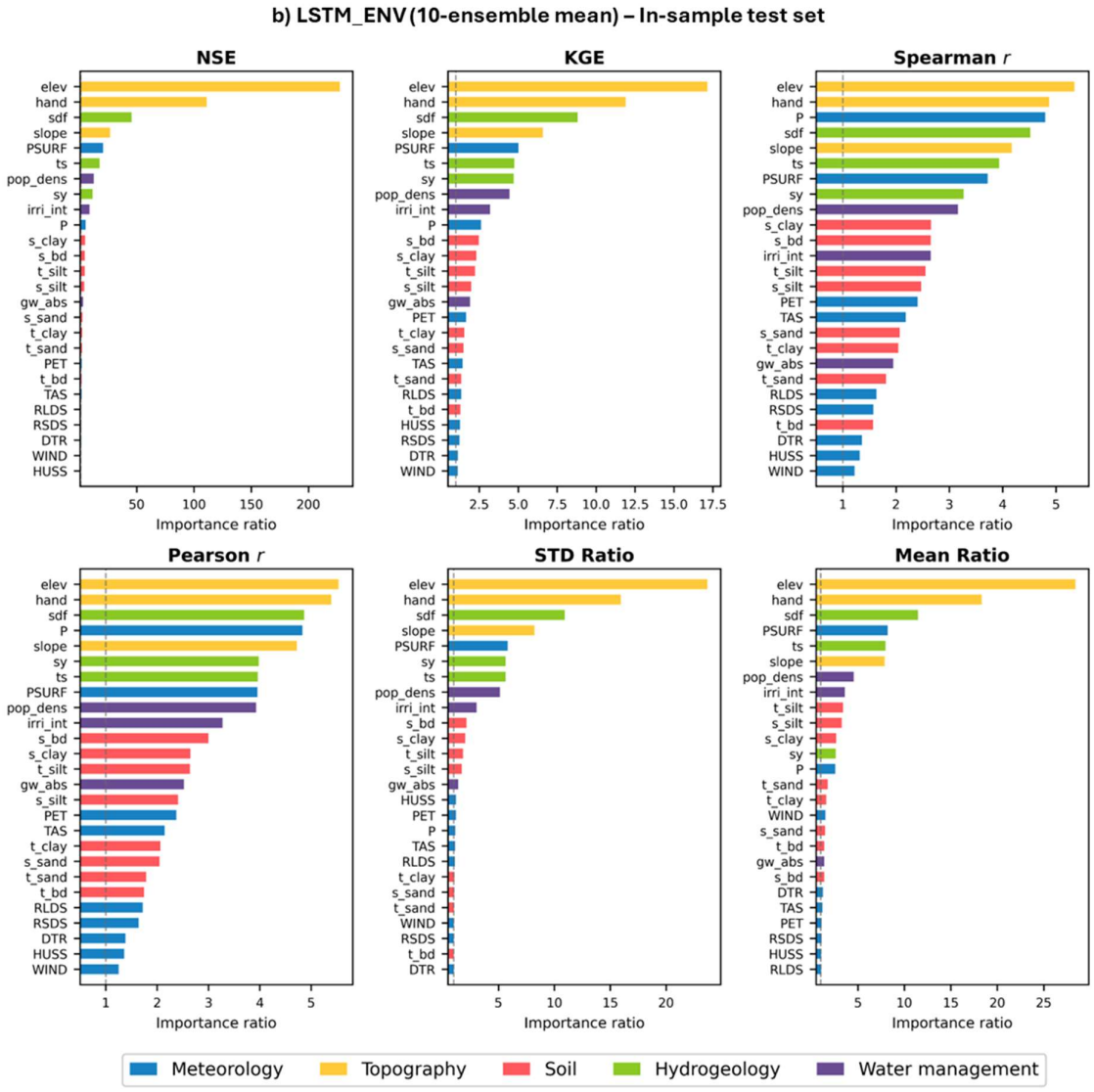


Fig. S12. Correlation between features (i.e. dynamic inputs averaged over 1961–2019, UPPERCASE, and static inputs, lowercase). Pearson correlations are calculated based on 636 stations used for training and 341unseen stations during training (977 stations in total). Value close to 1 and -1 represent stronger positive (red) and negative (blue) relationships, respectively.

Fig. S13. (Figure across three pages) Individual feature importance for LSTM_ENV on (a) the training set, (b) the in-sample test set, and (c) the out-of-sample test set. Dynamic inputs are timeseries of meteorological variables (blue, UPPERCASE). Static inputs (lowercase) are classified into four categories: topography (yellow), soil (red), hydrogeology (green), and water management (purple).




b) LSTM_ENV (10-ensemble mean) – In-sample test set
NSE
KGE
Spearman r
Pearson r
STD Ratio
Mean Ratio
Importance ratio
Meteorology
Topography
Soil
Hydrogeology
Water management

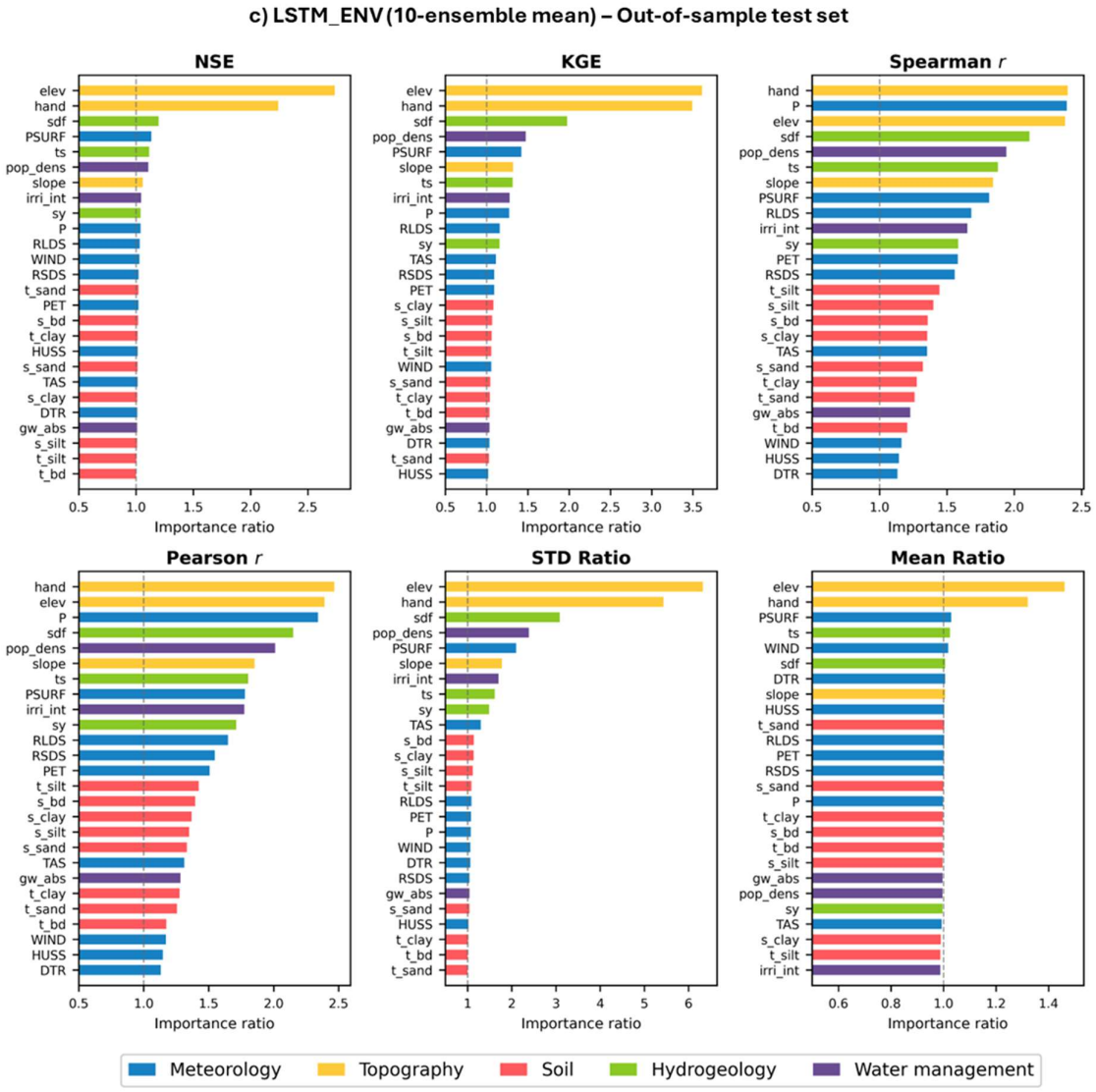
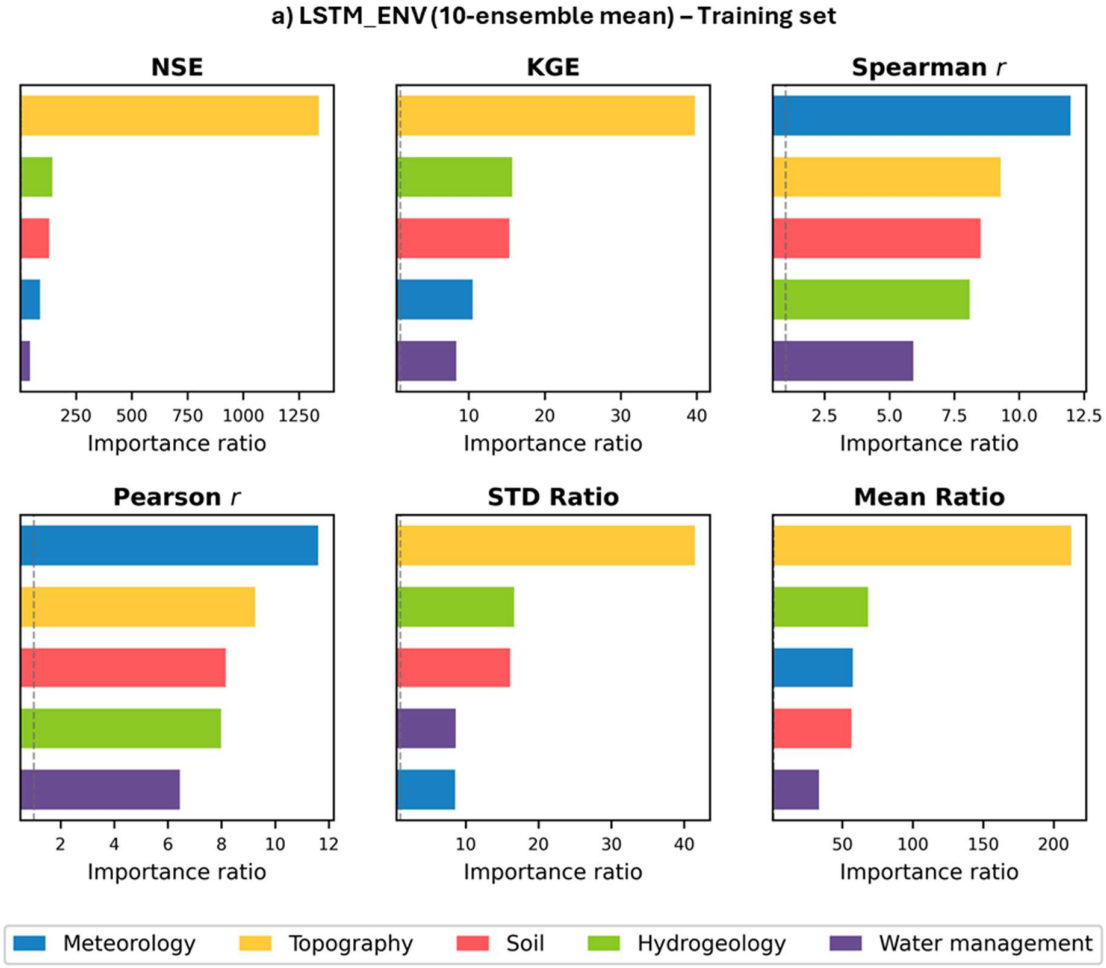
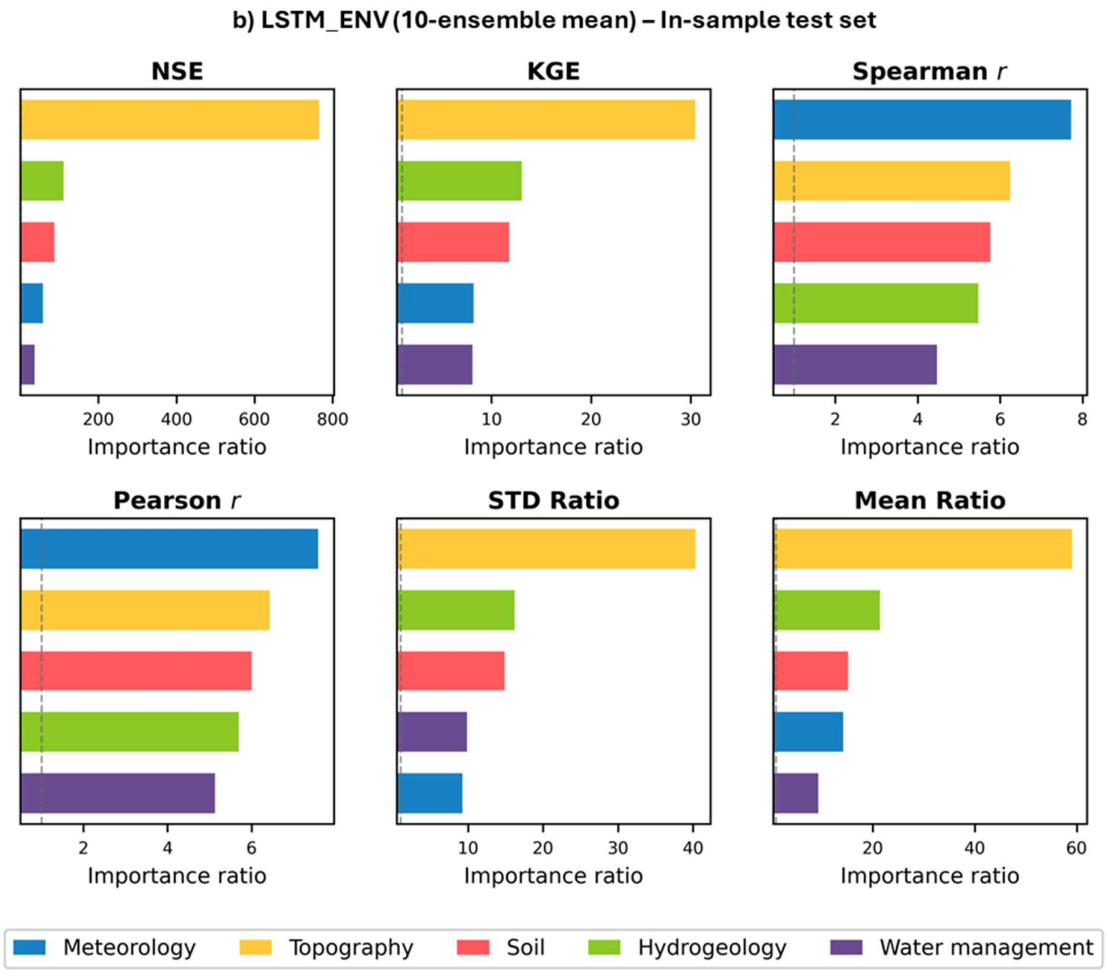

c) LSTM_ENV (10-ensemble mean) – Out-of-sample test set
NSE
elev
hand
sdf
PSURF
ts
pop_dens
slope
irri_int
sy
P
RLDS
WIND
RSDS
t_sand
PET
s_bd
t_clay
HUSS
s_sand
TAS
s_clay
DTR
gw_abs
s_silt
t_silt
t_bd
0.5 1.0 1.5 2.0 2.5
Importance ratio
KGE
elev
hand
sdf
pop_dens
PSURF
slope
ts
irri_int
P
RLDS
sy
TAS
RSDS
PET
s_clay
s_silt
s_bd
t_silt
WIND
s_sand
t_clay
t_bd
gw_abs
DTR
t_sand
HUSS
0.5 1.0 1.5 2.0 2.5 3.0 3.5
Importance ratio
Spearman r
hand
P
elev
sdf
pop_dens
ts
slope
PSURF
RLDS
irri_int
sy
PET
RSDS
t_silt
s_silt
s_bd
s_clay
TAS
s_sand
t_clay
t_sand
gw_abs
t_bd
WIND
HUSS
DTR
0.5 1.0 1.5 2.0 2.5
Importance ratio
Pearson r
hand
elev
P
sdf
pop_dens
slope
ts
PSURF
irri_int
sy
RLDS
RSDS
PET
t_silt
s_bd
s_clay
s_silt
s_sand
TAS
gw_abs
t_clay
t_sand
t_bd
WIND
HUSS
DTR
0.5 1.0 1.5 2.0 2.5
Importance ratio
STD Ratio
elev
hand
sdf
pop_dens
PSURF
slope
irri_int
ts
sy
TAS
s_bd
s_clay
s_silt
t_silt
RLDS
PET
P
WIND
DTR
RSDS
gw_abs
s_sand
HUSS
t_clay
t_bd
t_sand
1 2 3 4 5 6
Importance ratio
Mean Ratio
elev
hand
PSURF
ts
WIND
sdf
DTR
slope
HUSS
t_sand
RLDS
PET
RSDS
s_sand
P
t_clay
s_bd
t_bd
s_silt
gw_abs
pop_dens
sy
TAS
s_clay
t_silt
irri_int
0.6 0.8 1.0 1.2 1.4
Importance ratio
Meteorology
Topography
Soil
Hydrogeology
Water management

Fig. S14. (Figure across three pages) Grouped feature importance for LSTM_ENV on (a) the training set, (b) the in-sample test set, and (c) the out-of-sample test set. Dynamic inputs are timeseries of meteorological variables (blue, UPPERCASE). Static inputs (lowercase) are classified into four categories: topography (yellow), soil (red), hydrogeology (green), and water management (purple).




b) LSTM_ENV (10-ensemble mean) – In-sample test set
NSE
KGE
Spearman r
Pearson r
STD Ratio
Mean Ratio
Importance ratio
Meteorology
Topography
Soil
Hydrogeology
Water management

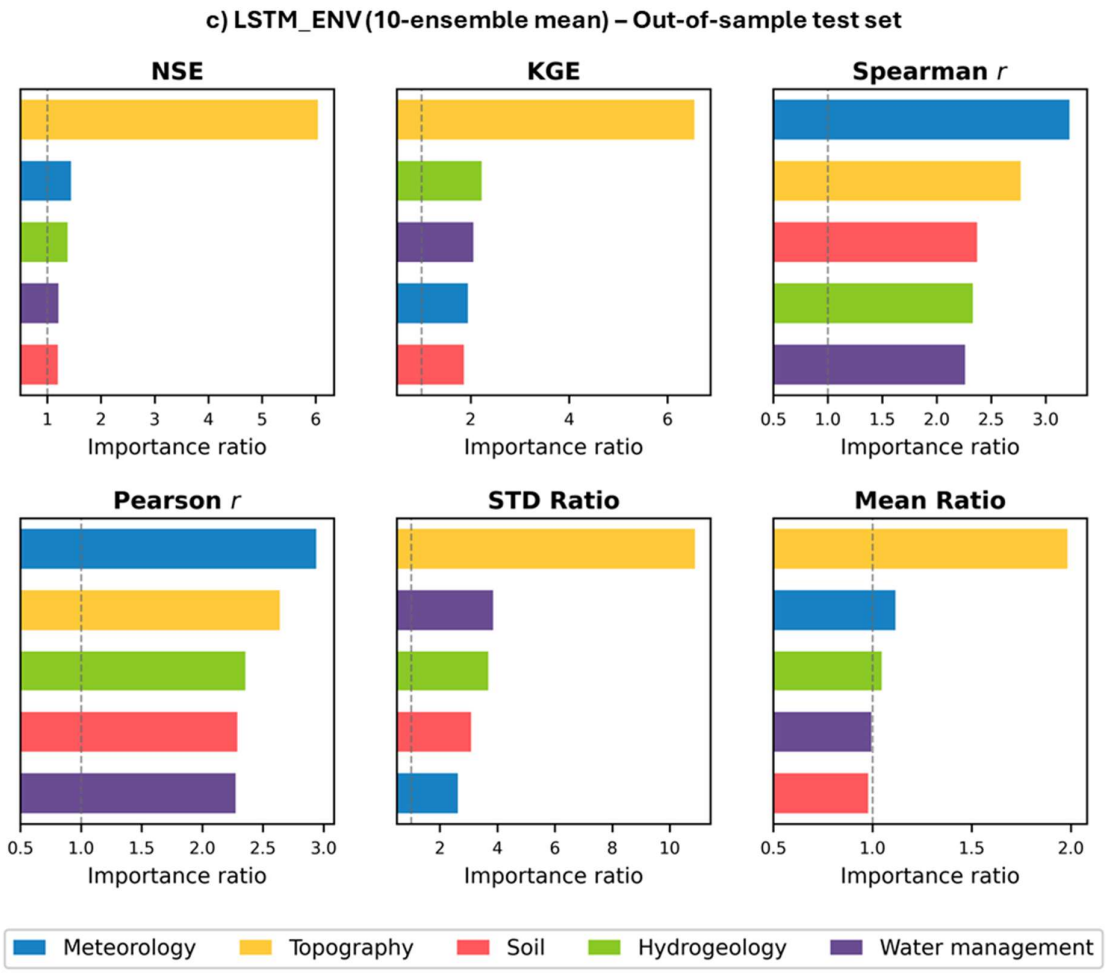
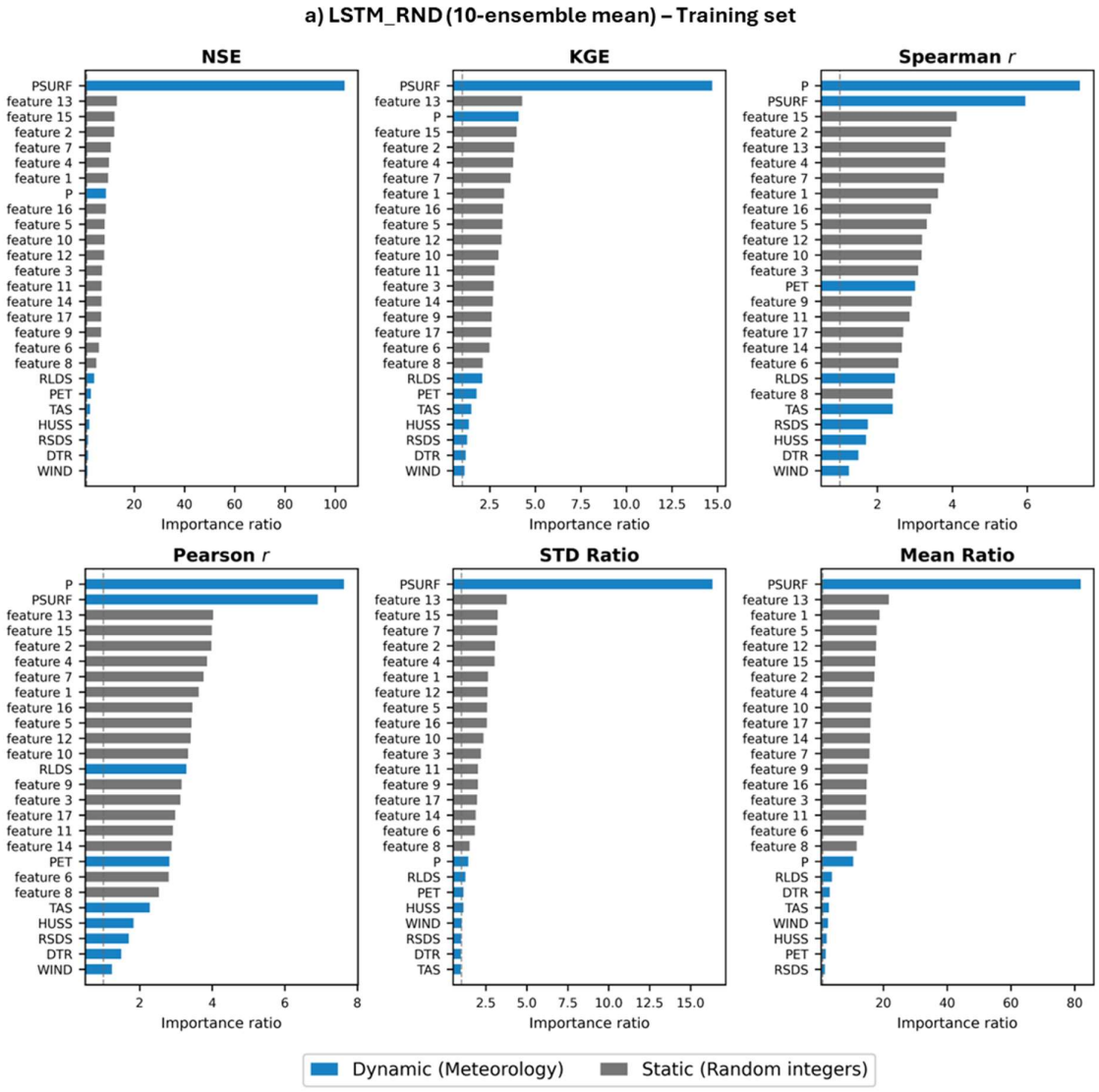
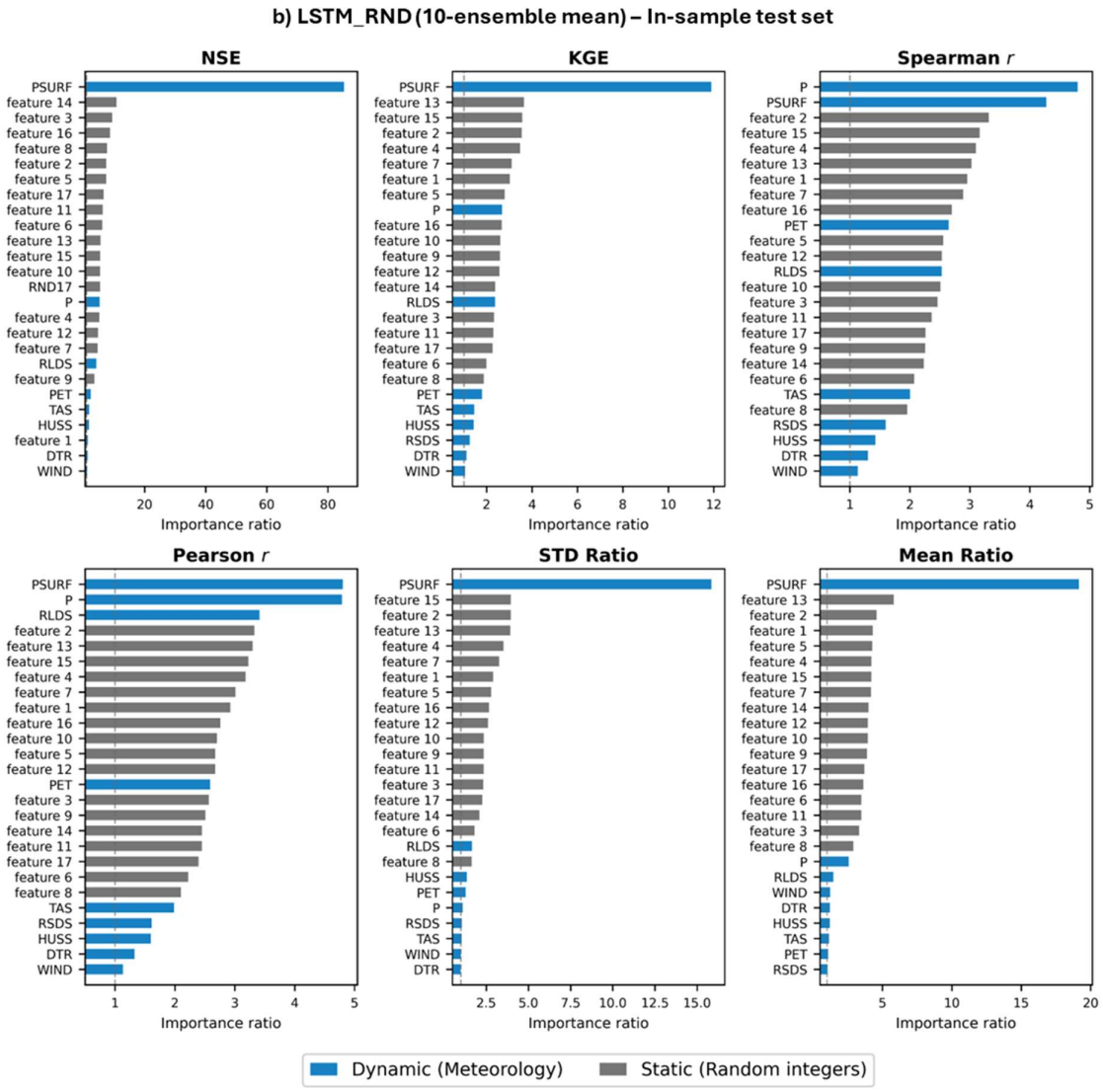
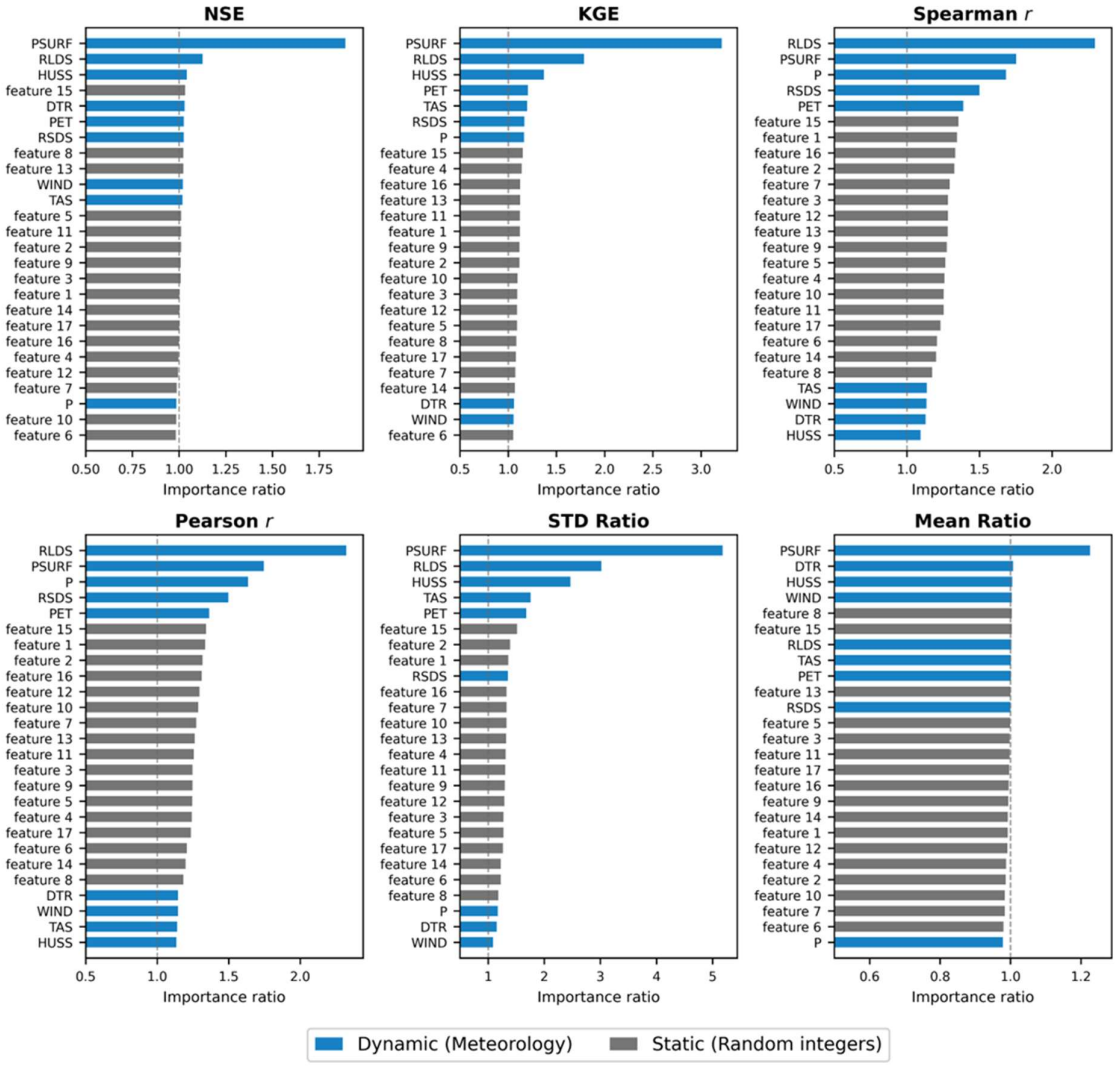

c) LSTM_ENV (10-ensemble mean) – Out-of-sample test set
NSE
KGE
Spearman r
Pearson r
STD Ratio
Mean Ratio
Importance ratio
Meteorology
Topography
Soil
Hydrogeology
Water management

Fig. S15. (Figure across three pages) Individual feature importance for LSTM_RND on (a) the training set, (b) the in-sample test set, and (c) the out-of-sample test set. Dynamic inputs are timeseries of meteorological variables (blue, UPPERCASE). Static inputs (grey) are random integers.




b) LSTM_RND (10-ensemble mean) – In-sample test set
NSE
PSURF
feature 14
feature 3
feature 16
feature 8
feature 2
feature 5
feature 17
feature 11
feature 6
feature 13
feature 15
feature 10
RND17
P
feature 4
feature 12
feature 7
RLDS
feature 9
PET
TAS
HUSS
feature 1
DTR
WIND
20 40 60 80
Importance ratio
KGE
PSURF
feature 13
feature 15
feature 2
feature 4
feature 7
feature 1
feature 5
P
feature 16
feature 10
feature 9
feature 12
feature 14
RLDS
feature 3
feature 11
feature 17
feature 6
feature 8
PET
TAS
HUSS
RSDS
DTR
WIND
2 4 6 8 10 12
Importance ratio
Spearman r
P
PSURF
feature 2
feature 15
feature 4
feature 13
feature 1
feature 7
feature 16
PET
feature 5
feature 12
RLDS
feature 10
feature 3
feature 11
feature 17
feature 9
feature 14
feature 6
TAS
feature 8
RSDS
HUSS
DTR
WIND
1 2 3 4 5
Importance ratio
Pearson r
PSURF
P
RLDS
feature 2
feature 13
feature 15
feature 4
feature 7
feature 1
feature 16
feature 10
feature 5
feature 12
PET
feature 3
feature 9
feature 14
feature 11
feature 17
feature 6
feature 8
TAS
RSDS
HUSS
DTR
WIND
1 2 3 4 5
Importance ratio
STD Ratio
PSURF
feature 15
feature 2
feature 13
feature 4
feature 7
feature 1
feature 5
feature 16
feature 12
feature 10
feature 9
feature 11
feature 3
feature 17
feature 14
feature 6
RLDS
feature 8
HUSS
PET
P
RSDS
TAS
WIND
DTR
2.5 5.0 7.5 10.0 12.5 15.0
Importance ratio
Mean Ratio
PSURF
feature 13
feature 2
feature 1
feature 5
feature 4
feature 15
feature 7
feature 14
feature 12
feature 10
feature 9
feature 17
feature 16
feature 6
feature 11
feature 3
feature 8
P
RLDS
WIND
DTR
HUSS
TAS
PET
RSDS
5 10 15 20
Importance ratio
Dynamic (Meteorology)
Static (Random integers)


c) LSTM_RND (10-ensemble mean) – Out-of-sample test set
NSE
KGE
Spearman r
Pearson r
STD Ratio
Mean Ratio
Importance ratio
Dynamic (Meteorology)
Static (Random integers)