In this study the authors propose a novel deep learning architecture for ground water level (GWL) forecasting. In a broad and comprehensive empirical evaluation across multiple geographical, meteorological and other conditions the authors find that the proposed model performs better than a number of baseline models. The study investigates a number of very interesting angles on ML time series models for GWL forecasting. Most importantly probably the authors take a close look at how different predictive variables can be combined to improve GWL forecasts, such as human influences, meteorological forcings or geological features and how these influences vary over time, space and when partitioned into different meteorological regimes. The authors investigate and report a number of different metrics. The selection is very sensible and provides a good intuition across the different aspects of what the models capture. The evaluation and experimental setup is very well organised and conducted. Especially the analysis and discussion of the impact of various geological, hydrological and other factors on the forecasting quality is interesting for a broad community of researchers and beyond. In the evaluation, the authors cover a broad spectrum of geological, meteorological and societal conditions. Such a comprehensive account of GWL forecasting challenges is an important contribution. It would be great if the authors could share the data set for research purposes publicly? A number of ideas in the modelling seem novel and helpful. This includes for instance the combination of multiple loss functions, the ideas of the multiple encoders and the combination of data sources. In some cases (e.g. the loss functions or the model architecture), it could be helpful to investigate some more aspects that could clarify the benefits of those methodological contributions.

Response

We sincerely appreciate the reviewer for the highly encouraging assessment of our work, the positive remarks on our multi-source framework, and the thoughtful recommendations to further clarify our methodological contributions. In response to the suggestions and specific questions raised, we have thoroughly revised the manuscript and conducted extensive additional experiments. Furthermore, to clearly delineate the advantages of our proposed architecture and loss formulation, we have incorporated percentile-based statistics across all evaluation tables, implemented an independent Optuna-based hyperparameter optimization procedure for each baseline model to ensure rigorous and fair comparisons, performed comprehensive sensitivity analysis on the multi-objective loss weights configurations, and elaborated on the architectural distinctions. We have also clarified the strict out-of-sample data preprocessing protocols to eliminate risk of data leakage and updated the relevant visual presentations. To support research transparency and reproducibility, the source code of HHA-Net and all baseline models has been deposited and made publicly available on Zenodo (https://doi.org/10.5281/zenodo.21915502, https://doi.org/10.5281/zenodo.18130111, https://doi.org/10.5281/zenodo.19230562, https://doi.org/10.5281/zenodo.20774213). Detailed point-by-point responses to each specific comment and the corresponding sentences revisions are presented below.

Areas for improvement

Generally it would be great for all reported metrics to include percentiles (maybe 25%/50%/75%) rather than average values and standard deviations. In particular for the Ablation studies, but also at other places, for instance Table 2, it would be great to get a better understanding of the range of the metrics.

Response

We sincerely appreciate the reviewer for this valuable suggestion. To address this, we have added percentile-based statistics (Q_{25}/Q_{50}/Q_{75}) to all performance tables and the ablation experiments in addition to the originally reported mean values, each now including both mean-based and quantile-based results. We also refine the evaluation of model performance and ablation experiments throughout the manuscript based on the quantile distributions.

We have added or revised the sentences in the manuscript as follows:

"HHA-Net exhibits different performance across site types (Table 2). At urban sites, HHA-Net achieves the MAPE of 1.03%-3.34% and RMSE of 0.30-0.64, reaching its optimal performance in the NCP (R^2 = 0.984, NSE = 0.975). For natural sites, the model maintains MAPE within 2.50%-5.95%, RMSE between 0.31 and 0.46, and R^2 ranging from 0.676 to 0.943. In the YTMR, the interquartile MAPE range (1.29%-3.59%) closely brackets the mean, indicating relatively uniform performance across the region; whereas for natural sites in NCP the mean MAPE of 5.95% considerably exceeds both the median (1.18%) and the third quartile (1.55%), demonstrating that the elevated mean is driven by some sites with anomalously large errors."

"The optimal hyperparameter configurations and predictive performance for baseline models, obtained through the Optuna-based hyperparameter optimization procedure, are summarized in Table S4 and S5. With these configurations, most baselines achieve positive mean NSE values in the NCP and NJP (e.g., 0.930 for Transformer at NCP urban sites and 0.926 for ST-GNN at NJP natural sites), whereas low or negative mean NSE values are concentrated in the YTMR. Overall, HHA-Net outperforms baselines across most regions and site types, particularly in the YTMR and NCP. At agricultural sites, HHA-Net reduces prediction errors compared to the top baselines (e.g., MAPE of 1.10% vs. Transformer's 1.93% in NCP; 3.86% vs. MLP's 7.78% in YTMR). At YTMR natural sites, GNN (10.30%) and Transformer (7.18%) exhibit mean MAPEs roughly 3 to 4 times higher than those of HHA-Net (2.50%), with wider and more right-skewed quantile ranges (e.g., 2.54%-17.51% for GNN) compared to those of HHA-Net (1.29%-3.59%). At YTMR urban sites, HHA-Net similarly yields major improvements, with ST-GNN's MAPE (13.97%) exceeding four times that of HHA-Net (3.34%). An exception occurs in the NJP region, where ST-GNN demonstrates superior mean performance at agricultural and natural sites (MAPE = 3.63% and 2.02%, vs. 4.72% and 2.85% for HHA-Net), likely reflecting the densely distributed wells and regular groundwater extraction patterns in the NJP that create well-defined spatial associations effectively captured by graph

structures. At NJP urban sites, where ST-GNN achieves a lower median MAPE (1.30% vs. 2.18% for HHA-Net), HHA-Net's R^2 quantiles (Q_{25}-Q_{75} = 0.869-0.971) consistently exceed those of ST-GNN (0.845–0.963) across all three quartiles, indicating that HHA-Net provides a consistently more reliable overall fit."

"Given the inherent uncertainty in GWD arising from aquifer heterogeneity and spatial variability, we selected an additional 35 sites in the Lower Yangtze River Plain (LYRP), which is a region characterized by intensive human activities, high urbanization levels, and significant industrialization, to validate the model's generalizability and performance (Song et al. 2016). Site information and data are summarized in Table S8, with the corresponding model results presented in Table S9. The results demonstrate that HHA-Net consistently provides high-precision predictions across the three site types, achieving MAPE values ranging from 2.25% to 4.92% and R^2 values ranging from 0.704 to 0.961 (Figure S4). Quantile-based analysis corroborates this consistency, with median (Q50) MAPE values of 3.55%, 3.68%, and 2.22% for agricultural, natural, and urban sites, respectively, and relatively narrow interquartile ranges, further confirming that HHA-Net maintains stable and reliable performance in regions with intensive human disturbance."

"In the NCP, the spatial distribution of performance is influenced by the intensity of human activities, with agricultural sites demonstrating the highest GWD prediction accuracy. However, one natural site located at the boundary between agricultural and urban areas shows abnormal performance with MAPE reaching 24.98%, which substantially inflates the regional mean error; indeed, the mean MAPE for natural sites in the NCP (5.95%) markedly exceeds the corresponding median and third quartile (1.18% and 1.55%, respectively)."

"Removing the historical encoder results in RMSE increases of 2213.2%, 1490.2%, and 1398.5% in the YTMR, NCP, and NJP, respectively, indicating that historical GWD states serve as the foundation for accurate predictions. This finding is further corroborated by the quantile-based RMSE, with the median RMSE increasing by more than an order of magnitude and the interquartile range (Q_{25}–Q_{75}) also shifting substantially upward across all three regions."

Table S4. Performance comparison between HHA-Net and baseline models across agricultural, natural, and urban sites in three study regions based on quantiles (25%/50%/75%) metrics.

Area		MAPE(%)			RMSE			NSE			R^2		
	Agricultural	Q_{25}	Q_{50}	Q_{75}	Q_{25}	Q_{50}	Q_{75}	Q_{25}	Q_{50}	Q_{75}	Q_{25}	Q_{50}	Q_{75}
	GNN	1.829	7.529	15.857	0.660	0.836	1.326	-3.618	-1.277	0.559	0.024	0.208	0.718
	CNN_LSTM	2.491	8.248	12.478	0.761	1.075	1.486	-1.840	-0.852	0.146	0.169	0.370	0.541
	MLP	2.879	4.823	9.066	0.631	1.096	1.400	-1.324	-0.257	-0.023	0.175	0.394	0.720
	ST_GNN	3.125	7.527	19.864	0.852	1.100	1.615	-6.507	-1.079	0.152	0.038	0.200	0.400
	Transformer	1.937	3.715	11.974	0.904	1.040	1.234	-1.671	-0.143	0.217	0.221	0.416	0.728
	HHA_Net	**1.381**	**2.885**	**4.488**	**0.327**	**0.665**	**0.773**	**0.455**	**0.593**	**0.656**	**0.653**	**0.727**	**0.819**
	Natural												
	GNN	2.535	5.279	17.505	0.609	0.694	1.088	-8.460	-4.572	-2.143	0.066	0.197	0.380
	CNN_LSTM	2.788	5.769	9.540	0.453	0.582	0.819	-2.104	-0.885	0.057	0.118	0.264	0.363
YTMR	MLP	3.125	5.058	8.642	0.488	0.619	0.899	-2.575	-1.289	-0.105	0.124	0.295	0.439
	ST_GNN	4.620	5.889	12.599	0.554	0.812	1.098	-9.402	-1.086	-0.577	0.042	0.080	0.278
	Transformer	2.521	3.316	6.735	0.414	0.466	0.911	-3.974	-0.568	-0.168	0.104	0.279	0.622
	HHA_Net	**1.288**	**1.804**	**3.591**	**0.193**	**0.245**	**0.363**	**0.197**	**0.521**	**0.764**	**0.630**	**0.732**	**0.824**
	Urban												
	GNN	3.006	9.068	11.466	0.709	0.897	1.045	-9.830	0.103	0.361	0.203	0.420	0.639
	CNN_LSTM	4.468	9.411	10.395	0.794	0.962	1.095	-0.653	-0.400	0.065	0.087	0.210	0.393
	MLP	4.069	7.973	9.064	0.769	0.878	1.040	-0.361	-0.193	0.030	0.051	0.105	0.233
	ST_GNN	5.247	9.446	13.753	0.712	1.210	1.756	-1.322	-0.083	-0.071	0.003	0.006	0.048
	Transformer	2.924	6.429	6.826	0.537	0.725	0.811	-0.559	0.346	0.509	0.379	0.570	0.681
	HHA_Net	**1.776**	**3.188**	**4.569**	**0.355**	**0.417**	**0.458**	**0.166**	**0.691**	**0.836**	**0.666**	**0.729**	**0.851**
NCP	*Agricultural*												

	GNN	2.235	3.497	4.998	0.605	0.990	1.175	0.350	0.809	0.940	0.791	0.932	0.972
	CNN_LSTM	1.442	1.578	2.453	0.423	0.552	0.872	0.786	0.917	0.962	0.829	0.968	0.978
	MLP	2.085	2.941	4.872	0.774	0.855	0.942	0.660	0.787	0.895	0.752	0.916	0.951
	ST_GNN	1.171	2.048	3.219	0.331	0.629	0.985	0.670	0.945	0.965	0.912	0.967	0.987
	Transformer	0.986	1.563	2.294	0.312	0.449	0.584	0.842	0.938	0.984	0.934	0.970	0.989
	HHA_Net	**0.663**	**0.966**	**1.098**	**0.206**	**0.281**	**0.395**	**0.955**	**0.983**	**0.988**	**0.968**	**0.988**	**0.993**
	Natural												
	GNN	3.003	5.109	5.576	1.219	1.248	1.378	0.144	0.804	0.909	0.328	0.873	0.923
	CNN_LSTM	2.189	2.213	3.668	0.586	0.692	0.987	0.464	0.925	0.953	0.482	0.948	0.959
	MLP	2.025	4.503	7.233	0.957	1.093	1.572	0.490	0.497	0.882	0.498	0.737	0.944
	ST_GNN	1.811	3.665	4.257	0.478	0.988	1.021	0.511	0.861	0.939	0.695	0.918	0.952
	Transformer	1.762	1.851	4.151	0.526	0.950	1.112	0.504	**0.941**	0.948	0.703	**0.952**	0.952
	HHA_Net	**0.901**	**1.181**	**1.552**	**0.272**	**0.468**	**0.532**	**0.838**	0.920	**0.989**	**0.901**	0.943	**0.989**
	Urban												
	GNN	2.557	3.880	6.432	0.940	1.017	1.142	0.448	0.716	0.888	0.722	0.833	0.915
	CNN_LSTM	1.607	2.329	2.935	0.451	0.555	0.849	0.874	0.898	0.938	0.920	0.945	0.957
	MLP	2.185	3.604	5.999	0.885	1.023	1.068	0.445	0.639	0.881	0.762	0.911	0.925
	ST_GNN	2.197	2.959	3.435	0.643	0.747	0.859	0.805	0.830	0.873	0.940	0.952	0.967
	Transformer	1.305	1.649	2.071	0.372	0.435	0.557	0.933	0.942	0.953	0.949	0.963	0.972
	HHA_Net	**0.805**	**1.005**	**1.280**	**0.188**	**0.286**	**0.395**	**0.968**	**0.970**	**0.986**	**0.978**	**0.986**	**0.989**
NJP	*Agricultural*												
	GNN	1.090	2.939	9.255	0.250	0.302	0.421	0.420	0.631	0.949	0.629	0.727	0.951
	CNN_LSTM	1.528	4.005	10.010	0.265	0.299	0.498	0.344	0.694	0.925	0.618	0.826	0.967
	MLP	1.343	4.507	11.838	0.332	0.393	0.508	0.029	0.485	0.893	0.530	0.706	0.947
	ST_GNN	**0.556**	**1.497**	5.947	**0.138**	**0.189**	**0.224**	**0.821**	0.866	**0.978**	0.847	**0.927**	**0.981**
	Transformer	1.885	8.151	27.765	0.497	0.672	0.922	-2.675	-0.663	0.857	0.642	0.780	0.934

HHA_Net	1.876	***2.663***	**5.169**	***0.176***	0.300	0.426	***0.532***	**0.880**	***0.953***	**0.853**	***0.912***	0.965
Natural												
GNN	0.844	1.103	3.387	0.211	0.248	0.381	0.842	0.914	0.962	0.881	0.920	0.973
CNN_LSTM	1.209	1.928	3.276	0.288	0.418	0.535	0.767	0.833	0.920	0.862	0.919	0.948
MLP	1.319	2.380	3.679	0.323	0.444	0.537	0.807	0.868	0.915	0.866	0.915	0.943
ST_GNN	**0.586**	**0.727**	**1.537**	**0.144**	**0.202**	**0.243**	**0.933**	**0.977**	**0.985**	**0.947**	**0.987**	**0.992**
Transformer	1.476	1.992	7.333	0.335	0.628	0.829	0.737	0.801	0.932	0.841	0.960	0.974
HHA_Net	1.171	2.184	***2.576***	0.255	0.319	0.520	0.738	***0.927***	0.944	***0.913***	***0.968***	0.972
Urban												
GNN	1.355	1.828	5.250	0.262	0.419	0.722	0.632	0.733	0.889	0.657	0.810	0.895
CNN_LSTM	1.523	2.098	4.265	0.299	0.422	0.671	0.609	0.744	0.899	0.741	0.856	0.933
MLP	1.960	3.143	6.300	0.388	0.570	0.752	0.252	0.628	0.775	0.658	0.752	0.881
ST_GNN	**0.871**	**1.304**	**3.098**	**0.181**	**0.286**	0.521	**0.802**	**0.883**	0.959	0.845	0.908	0.963
Transformer	1.933	2.466	7.421	0.472	0.649	0.929	-0.243	0.581	0.779	0.684	0.805	0.881
HHA_Net	1.445	2.182	***3.325***	***0.225***	***0.400***	**0.464**	***0.751***	***0.854***	**0.960**	**0.869**	**0.941**	**0.971**

Table S9. Performance comparison between HHA-Net and baseline models across agricultural, natural, and urban sites in the Lower Yangtze River Plain (LYRP).

(a). Performance comparison based on mean metrics.

		Agricultural sites				Natural sites				Urban sites			
		MAPE(%)	RMSE	NSE	R^2	MAPE(%)	RMSE	NSE	R^2	MAPE(%)	RMSE	NSE	R^2
LYRP	GNN	7.433	0.389	-0.994	0.560	7.000	0.360	-0.860	0.680	2.511	0.268	0.878	0.913
	CNN_LSTM	5.679	0.254	-0.165	0.533	5.901	0.273	-1.852	0.515	3.746	0.327	0.250	0.876
	MLP	6.747	0.300	-0.594	0.527	6.348	0.316	-1.266	0.507	3.473	0.316	0.660	0.868
	ST_GNN	5.486	0.237	-0.560	0.683	5.129	0.249	0.053	0.695	2.580	0.245	0.869	0.924
	Transformer	5.626	**0.213**	-0.021	0.664	4.674	**0.213**	0.110	0.690	**1.943**	**0.184**	0.888	0.924
	HHA-Net	**4.923**	0.232	**0.330**	**0.704**	**3.559**	0.266	**0.619**	**0.858**	2.256	0.232	**0.919**	**0.961**

(b). Performance comparison based on quantiles (25%/50%/75%) metrics.

Area		MAPE(%)			RMSE			NSE			R^2		
	Agricultural	Q_{25}	Q_{50}	Q_{75}	Q_{25}	Q_{50}	Q_{75}	Q_{25}	Q_{50}	Q_{75}	Q_{25}	Q_{50}	Q_{75}
	GNN	3.363	6.885	8.832	0.284	0.343	0.473	-2.319	0.164	0.949	0.206	0.522	0.989
	CNN_LSTM	2.020	4.054	6.156	0.209	0.249	0.291	-1.484	0.599	0.970	0.040	0.619	0.982
	MLP	2.359	4.458	8.572	0.266	0.302	0.343	-1.697	0.249	0.940	0.087	0.555	0.951
	ST_GNN	1.544	**3.032**	7.796	0.170	0.218	0.275	-0.515	**0.666**	0.981	0.422	0.731	0.986
	Transformer	**1.101**	3.586	6.767	0.185	0.200	**0.248**	-1.079	0.423	**0.984**	0.327	0.691	**0.994**
	HHA_Net	2.313	3.554	**4.327**	**0.163**	**0.193**	0.257	0.120	0.636	0.970	**0.443**	**0.791**	0.993
	Natural												
	GNN	5.502	5.708	9.357	0.270	0.297	0.329	-1.912	0.268	0.313	0.603	0.679	0.799
	CNN_LSTM	**3.381**	5.064	6.601	0.167	0.285	0.333	-1.319	0.342	0.772	0.232	0.558	0.796
LYRP	MLP	4.182	5.660	6.294	0.218	0.263	0.357	-0.610	0.291	0.601	0.262	0.599	0.723
	ST_GNN	4.374	4.991	6.149	0.186	0.218	0.308	-0.129	0.485	0.595	0.672	0.756	0.842
	Transformer	3.416	4.320	5.945	0.176	0.207	**0.235**	0.030	0.556	0.700	0.618	0.769	0.855
	HHA_Net	3.568	**3.680**	**4.001**	**0.113**	**0.163**	0.250	**0.524**	**0.680**	**0.802**	**0.824**	**0.847**	**0.920**
	Urban												
	GNN	1.640	2.195	3.137	0.185	0.196	0.249	0.914	0.914	0.967	0.943	0.975	0.985
	CNN_LSTM	2.144	2.621	3.932	0.258	0.306	0.369	0.775	0.775	0.963	0.874	0.976	0.986
	MLP	2.333	2.755	3.492	0.253	0.311	0.365	0.836	0.836	0.956	0.865	0.950	0.976
	ST_GNN	2.038	2.368	2.758	0.202	0.252	0.290	0.905	0.905	0.969	0.934	0.976	0.988
	Transformer	**1.162**	**1.358**	**1.920**	**0.131**	**0.167**	**0.219**	**0.943**	0.943	**0.989**	0.966	0.985	0.992
	HHA_Net	1.788	2.224	2.634	0.147	0.222	0.291	0.872	**0.959**	0.978	**0.967**	**0.988**	**0.993**

Table S10. Performance comparison of the HHA-Net and baseline models (CNN and LSTM) on the benchmark dataset GEMS-GER (test dataset) including mean and quantiles (25%/50%/75%) metrics.

MAE	RMSE	NSE	R^2

	Mean	Q_{25}	Q_{50}	Q_{75}	Mean	Q_{25}	Q_{50}	Q_{75}	Mean	Q_{25}	Q_{50}	Q_{75}	Mean	Q_{25}	Q_{50}	Q_{75}
CNN	0.290	0.065	0.134	0.273	0.329	0.082	0.167	0.320	-2.299	0.212	0.763	0.915	0.834	0.778	0.908	0.958
LSTM	0.235	**0.059**	0.110	0.215	0.268	**0.075**	0.138	0.262	-1.806	0.431	0.826	0.938	0.877	0.849	0.941	0.974
HHA-Net	**0.150**	0.061	**0.096**	**0.164**	**0.186**	0.079	**0.122**	**0.201**	**0.348**	**0.633**	**0.863**	**0.940**	**0.901**	**0.882**	**0.954**	**0.979**

Table S11. Results of ablation experiments at the encoder level and mechanism level.
(a). Results of ablation experiments based on mean metrics.

	YTMR		NCP		NJP	
	RMSE	Impact	RMSE	Impact	RMSE	Impact
Full Model	0.0394	0	0.0700	0	0.0670	0
Hist Encoder	0.9118	2213.2%	1.1126	1490.2%	1.0048	1398.5%
Met Encoder	0.0603	53.0%	0.1523	117.1%	0.1724	157.3%
Geo Encoder	0.2685	581.2%	0.1835	162.3%	0.0750	11.8%
Human Encoder	0.1418	259.7%	0.1823	160.4%	0.0914	36.4%
Spatial Attention	1.0609	2591.3%	0.8367	1095.8%	1.0198	1420.9%
Site-type Awareness	2.0512	5103.7%	1.7744	2436.1%	0.9990	1390.0%
CNN Branch	0.6856	1639.4%	0.1383	97.6%	0.1326	97.7%
LSTM Branch	0.0587	48.9%	0.5026	618.4%	0.4293	540.3%

(b). Results of ablation experiments based on quantiles (25%/50%/75%) metrics.

	YTMR			NCP			NJP		
	Q_{25}	Q_{50}	Q_{75}	Q_{25}	Q_{50}	Q_{75}	Q_{25}	Q_{50}	Q_{75}
Full Model	0.0053	0.0140	0.0325	0.0214	0.0410	0.0707	0.0207	0.0373	0.0663
Hist Encoder	0.0487	0.2223	1.1788	0.4629	0.9511	1.3825	0.1528	0.7460	1.2809
Met Encoder	0.0113	0.0235	0.0504	0.0377	0.0916	0.1763	0.0466	0.0893	0.1677
Geo Encoder	0.0171	0.0425	0.1136	0.0224	0.0488	0.0833	0.0206	0.0418	0.0774
Human Encoder	0.0131	0.0279	0.0662	0.0578	0.1125	0.2153	0.0315	0.0563	0.0918
Spatial Attention	0.0482	0.2934	1.2453	0.3627	0.7329	1.0727	0.1261	0.6320	1.0814
Site-type Awareness	1.0664	1.5182	2.6283	0.9312	1.6456	2.3007	0.4886	0.7260	1.1208
CNN Branch	0.0643	0.1439	0.3708	0.0437	0.0774	0.1091	0.0421	0.0722	0.1140
LSTM Branch	0.0085	0.0242	0.0517	0.3068	0.4988	0.5899	0.1339	0.2480	0.4780

Sometimes it can be difficult to avoid leakage from test data to validation/training data. Especially when data cleaning is performed or data preprocessing, such as imputation, normalisation and other measures. It would be good if the authors directly stated that none of the test data was used for any preprocessing steps of the data engineering. This seems like a minor detail, but actually getting the mean right for the test period (which would be a lot easier when using the test data in a normalisation step) can have a large impact on the models, as the authors nicely show in the ablation studies on the temporal encoder.

Response

We sincerely appreciate the reviewer for pointing out this critical issue regarding rigorous machine learning evaluation. We would like to clarify that our data preprocessing strictly adheres to out-of-sample evaluation principles to prevent any form of data leakage. The continuous temporal dataset was strictly divided into training, validation, and test sets along the time sequence. All feature scaling and normalization procedures for both input features and target predictions were fitted strictly on the training split. The mean and standard deviation parameters derived from the training set were then applied to transform the validation and test sets. We have added the corresponding description in the manuscript as follows:

"To strictly prevent data leakage from the validation and test sets during data preprocessing, the parameters used to normalize input sequences and target variables were fitted exclusively on the training set. The fitted parameters were subsequently applied to transform the validation and test sets out of sample. No statistical information or global parameters from the evaluation periods were utilized in any data scaling or imputation procedures."

The results in Table 2 could be discussed in more detail with respect to the surprisingly low NSE for all baseline models in YTMR and 4 out of 6 models in NJP. This also relates to the point on the hyperparameter optimization below; it seems like with a bit more optimization (of hyperparameters or the model parameters themselves) for the other models, they would probably reach better NSEs.

Response

We sincerely appreciate the reviewer for this insightful comment. To address it, we performed an independent hyperparameter optimization for each baseline model in each region. This optimization substantially improved the performance of the baseline models, and most tuned baselines now achieve positive mean NSE values in the NCP and NJP, for example 0.930 for the Transformer at NCP urban sites and 0.926 for ST-GNN at NJP natural sites, whereas the remaining low or negative mean NSE values are concentrated in the YTMR. To clarify these remaining low values, we added discussion based on quantile metrics (Q_{25}, Q_{50}, and Q_{75}) for all models and regions. As NSE is bounded above at 1 but unbounded below, a few extreme outlying sites can severely deflate regional arithmetic means. For instance, the mean NSE of CNN-LSTM at YTMR natural sites (-145.23) lies almost two orders of magnitude below its own first quartile (-2.10), despite a median NSE of -0.89. Moreover, the quantile metrics indicates the real performance differences between the baseline models and the HHA-Net. All five baseline models still show negative median NSE values in the YTMR (-4.57 to -0.14) and interquartile MAPE ranges spanning more than sixfold, whereas HHA-Net maintains positive NSE values across its entire interquartile range with consistently narrow MAPE spans.

We have revised the corresponding sentences in the manuscript as follows:

"The optimal hyperparameter configurations and predictive performance for baseline models, obtained through the Optuna-based hyperparameter

optimization procedure, are summarized in Table S4 and S5. With these configurations, most baselines achieve positive mean NSE values in the NCP and NJP (e.g., 0.930 for Transformer at NCP urban sites and 0.926 for ST-GNN at NJP natural sites), whereas low or negative mean NSE values are concentrated in the YTMR. Overall, HHA-Net outperforms baselines across most regions and site types, particularly in the YTMR and NCP. At agricultural sites, HHA-Net reduces prediction errors compared to the top baselines (e.g., MAPE of 1.10% vs. Transformer's 1.93% in NCP; 3.86% vs. MLP's 7.78% in YTMR). At YTMR natural sites, GNN (10.30%) and Transformer (7.18%) exhibit mean MAPEs roughly 3 to 4 times higher than those of HHA-Net (2.50%), with wider and more right-skewed quantile ranges (e.g., 2.54%-17.51% for GNN) compared to those of HHA-Net (1.29%-3.59%). At YTMR urban sites, HHA-Net similarly yields major improvements, with ST-GNN's MAPE (13.97%) exceeding four times that of HHA-Net (3.34%). An exception occurs in the NJP region, where ST-GNN demonstrates superior mean performance at agricultural and natural sites (MAPE = 3.63% and 2.02%, vs. 4.72% and 2.85% for HHA-Net), likely reflecting the densely distributed wells and regular groundwater extraction patterns in the NJP that create well-defined spatial associations effectively captured by graph structures. At NJP urban sites, where ST-GNN achieves a lower median MAPE (1.30% vs. 2.18% for HHA-Net), HHA-Net's R^2 quantiles (Q_{25}-Q_{75} = 0.869-0.971) consistently exceed those of ST-GNN (0.845–0.963) across all three quartiles, indicating that HHA-Net provides a consistently more reliable overall fit.

Because NSE is unbounded below, a few extreme outlying sites can severely deflate regional mean NSEs, explaining the sharp discrepancies between low or negative means and vastly superior medians observed across baselines (Knoben et al. 2019, McCuen et al. 2006). For example, CNN-LSTM's mean NSE at YTMR natural sites (-145.23) lies almost two orders of magnitude below its first quartile (-2.10). Similarly, the mean NSEs of ST-GNN (-22.21) and GNN (-11.23) at YTMR agricultural sites are far lower than their medians (-1.08 and -1.28). In the NJP, Transformer's near-zero mean NSE (0.04) is likewise skewed by outlying sites, as its median NSE reaches 0.80 and its first quartile (0.74) lies well above the mean. Even in terms of medians, baseline NSEs across the YTMR remain negative (-4.57 to -0.14). Furthermore, the interquartile MAPEs of ST-GNN (3.13%-19.86%) and Transformer (1.94%-11.97%) each span more than a sixfold range at YTMR agricultural sites, while Transformer's interquartile MAPE similarly spans 1.89%-27.77% at NJP agricultural sites. In contrast, HHA-Net maintains positive NSEs across its entire interquartile range (Q_{25} = 0.455 and 0.197 at YTMR agricultural and natural sites) and consistently narrow MAPE spans. These comparisons demonstrate that guidance from hydrological domain knowledge significantly enhances model robustness and stability."

Table 2. Performance comparison between HHA-Net and baseline models across agricultural, natural, and urban sites in three study regions.

		Agricultural sites				Natural sites				Urban sites			
		MAPE(%)	RMSE	NSE	R^2	MAPE(%)	RMSE	NSE	R^2	MAPE(%)	RMSE	NSE	R^2
YTMR	GNN	9.808	1.827	-11.230	0.380	10.298	1.100	-8.244	0.253	9.791	1.304	-6.515	0.437
	CNN_LSTM	9.454	1.544	-16.394	0.363	7.390	1.574	-145.233	0.259	7.843	0.999	-0.488	0.255
	MLP	7.780	1.263	-5.034	0.427	6.549	1.029	-24.622	0.285	6.622	1.045	-2.127	0.201
	ST_GNN	10.89	1.781	-22.205	0.242	8.898	1.239	-13.848	0.180	13.967	1.857	-10.805	0.035
	Transformer	8.526	1.010	-1.570	0.454	7.183	0.715	-3.568	0.373	5.628	0.990	-2.431	0.500
	HHA_Net	**3.855**	**0.611**	**0.437**	**0.714**	**2.503**	**0.306**	**0.336**	**0.676**	**3.336**	**0.639**	**0.038**	**0.725**
NCP	GNN	3.979	0.941	0.599	0.856	17.677	1.321	-1.647	0.643	4.640	1.029	0.631	0.760
	CNN_LSTM	2.801	0.676	0.801	0.877	6.316	0.698	0.660	0.754	2.462	0.650	0.892	0.923
	MLP	3.940	0.938	0.736	0.840	12.752	1.189	-0.467	0.710	3.886	0.950	0.647	0.853
	ST_GNN	2.913	0.696	0.797	0.907	**5.004**	0.776	**0.744**	0.813	2.856	0.750	0.848	0.938
	Transformer	1.931	0.496	0.897	0.929	11.304	0.874	-0.194	0.821	1.808	0.488	0.930	0.946
	HHA_Net	**1.097**	**0.296**	**0.966**	**0.978**	***5.947***	**0.459**	***0.691***	**0.943**	**1.027**	**0.302**	**0.975**	**0.984**
NJP	GNN	5.860	0.343	0.606	0.753	2.881	0.403	0.859	0.883	3.819	0.525	0.433	0.736
	CNN_LSTM	7.104	0.374	0.529	0.775	3.573	0.447	0.782	0.862	4.014	0.511	0.476	0.775
	MLP	9.341	0.455	0.141	0.718	3.530	0.475	0.747	0.832	4.473	0.648	0.471	0.712
	ST_GNN	**3.628**	**0.203**	**0.868**	**0.898**	**2.016**	**0.232**	**0.926**	**0.954**	2.601	**0.357**	**0.716**	0.857
	Transformer	16.942	0.713	-1.458	0.774	6.939	0.601	0.040	0.892	6.869	0.730	-0.122	0.738
	HHA_Net	***4.722***	***0.332***	***0.733***	***0.883***	***2.854***	0.432	0.799	***0.923***	**2.525**	***0.394***	***0.525***	**0.892**

For the comparisons with the baseline models it would be helpful to elaborate a bit more on the hyper parameters. The authors write that "All baseline models adopt identical training strategies and hyperparameter settings as HHA-Net to ensure fair comparative experiments." - but usually different models require different hyper parameters. It is good practice to conduct some sort of hyper parameter optimization to ensure fair comparisons, rather than using the same hyper parameters for all. While it's true that often times hyper parameters have a limited impact on the outcome, in this case I would reckon that the individual architectures, dropout rates, optimization parameters could actually change the results.

Response

We sincerely appreciate the reviewer for this valuable and constructive comment. We fully agree that using identical hyperparameters across architecturally distinct models is not an ideal practice for ensuring fair comparison. To address this concern, we have replaced the original uniform-hyperparameter setting with an independent hyperparameter optimization (HPO) procedure for every baseline model (MLP, CNN-LSTM, Transformer, GNN, ST-GNN) in each study region. Specifically, we implemented an Optuna-based HPO pipeline using a Tree-structured Parzen Estimator sampler with a fixed random seed. For each model, we defined a search space covering the dropout rate, learning rate, and model-specific architectural parameters (e.g., hidden-layer size for CNN-LSTM/GNN/ST-GNN, hidden-layer configuration for MLP, and embedding dimension/number of heads/number of layers for the Transformer). The best-performing hyperparameter configuration for each model was then used to train the baseline models using the same data preprocessing procedure as HHA-Net, so that the only source of performance difference is the model architecture itself.

The resulting optimal hyperparameter configurations for all five baseline models in the study regions are now reported in Table S5. We have updated Section 2.2 (Benchmark Models) to describe this optimization procedure in detail, and we have re-run all baseline model evaluations using these individually optimized hyperparameters; the corresponding results, tables and figures throughout the manuscript have been updated accordingly. Notably, the individual architectures, dropout rates and hyperparameter optimization led to improvements in baseline model performance across all study regions. The code of the baseline models (run_models.py) has been made publicly available on Zenodo (https://doi.org/10.5281/zenodo.21915502).

We have revised the sentences in the manuscript as follows:

"This study selects five representative baseline models for comparison with the novel HHA-Net model, including Multi-Layer Perceptron (MLP), Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM), Transformer, Graph Neural Networks (GNN) and Spatio-Temporal Graph Neural Networks (ST-GNN). The MLP represents a feedforward neural network architecture that learns complex relationships between inputs and outputs through nonlinear transformations across multiple fully connected layers (Riedmiller and Lernen 2014). The MLP model flattens time series data and employs three hidden layers for GWD prediction, with each layer followed by a ReLU activation function for

nonlinear transformation and a dropout layer to prevent overfitting. The output layer contains fourteen neurons to generate fourteen-step predictions.

The CNN-LSTM represents a hybrid neural network architecture that combines the local feature extraction capabilities of CNNs with the sequential modeling advantages of LSTMs for enhanced time series forecasting (Zhao et al. 2025). The CNN-LSTM model employs a hierarchical processing strategy with two one-dimensional convolutional layers for spatiotemporal feature extraction, each followed by a ReLU activation function and a dropout layer. The convolutional outputs are reshaped before being fed into the LSTM layer, which models long-term temporal dependencies in the sequence.

The Transformer represents a neural network architecture that relies entirely on attention mechanisms to model sequence dependencies, enabling simultaneous focus on relationships between any two positions without distance limitations (Han et al. 2021). The Transformer model employs an encoder architecture with multiple stacked layers and multi-head self-attention for time series prediction, utilizing sine and cosine positional encoding to capture temporal information. Each encoder layer incorporates multi-head self-attention mechanisms and feedforward networks with residual connections, layer normalization, and dropout for training stability, with the final output layer generating 14-dimensional predictions for fourteen-step GWD prediction.

GNNs are designed to overcome the limitations of traditional neural networks, which are primarily built for data with fixed, regular structures (e.g., grids or sequences) and thus cannot directly operate on graph-structured data with irregular topological connectivity. Their core mechanism enables each node to update its representation by aggregating feature information from neighboring nodes as defined by the graph structure (Asif et al. 2021, Bai and Tahmasebi 2023). This process operates layer by layer, enabling the model to capture both local structural characteristics and broader global information, thereby effectively extracting feature representations at the nodes, edges, and entire graph level. The GNN model constructed in this study compresses node features into a fixed-dimensional vector through global average pooling following graph convolution computation, then sequentially passes them through a linear layer, a ReLU activation function, and a dropout layer, before finally mapping to 14-dimensional predictions via the output layer.

Building upon GNNs, ST-GNNs can simultaneously model spatial dependencies and temporal dynamics (Martinez-Ruiz et al. 2025, Taccari et al. 2024). ST-GNNs typically employ a hybrid architecture where GNN layers capture spatial topological structures while LSTMs or Temporal Convolutional Networks (TCNs) characterize temporal evolution patterns. This dual modeling capability enables the model to understand both spatial interactions between nodes and how these interactions evolve over time, demonstrating superior performance in complex dynamic system prediction. The key distinction of the ST-GNN model lies in the feature fusion stage, where extracted temporal features are concatenated with spatial features to form a hybrid vector, which is then processed through a fully

connected layer, a ReLU activation function, and a dropout layer before the output layer generates 14-dimensional predictions.

To ensure a rigorous and fair comparison, an independent hyperparameter optimization procedure was conducted for each baseline model using the Optuna framework with a Tree-structured Parzen Estimator sampler. For each model, the search space included the dropout rate (0.1-0.5), the learning rate (1×10^{-4}-1×10^{-2}, log-uniform), and model-specific architectural hyperparameters (e.g., the hidden-layer dimension for CNN-LSTM, GNN, and ST-GNN; the hidden-layer configuration for MLP; and the embedding dimension, number of attention heads, and number of encoder layers for the Transformer). The resulting optimal configuration for each baseline model was then used to retrain the final model following the same data preprocessing procedure as HHA-Net."

"The optimal hyperparameter configurations and predictive performance for baseline models, obtained through the Optuna-based hyperparameter optimization procedure, are summarized in Table S4 and S5. With these configurations, most baselines achieve positive mean NSE values in the NCP and NJP (e.g., 0.930 for Transformer at NCP urban sites and 0.926 for ST-GNN at NJP natural sites), whereas low or negative mean NSE values are concentrated in the YTMR. Overall, HHA-Net outperforms baselines across most regions and site types, particularly in the YTMR and NCP. At agricultural sites, HHA-Net reduces prediction errors compared to the top baselines (e.g., MAPE of 1.10% vs. Transformer's 1.93% in NCP; 3.86% vs. MLP's 7.78% in YTMR). At YTMR natural sites, GNN (10.30%) and Transformer (7.18%) exhibit mean MAPEs roughly 3 to 4 times higher than those of HHA-Net (2.50%), with wider and more right-skewed quantile ranges (e.g., 2.54%-17.51% for GNN) compared to those of HHA-Net (1.29%-3.59%). At YTMR urban sites, HHA-Net similarly yields major improvements, with ST-GNN's MAPE (13.97%) exceeding four times that of HHA-Net (3.34%). An exception occurs in the NJP region, where ST-GNN demonstrates superior mean performance at agricultural and natural sites (MAPE = 3.63% and 2.02%, vs. 4.72% and 2.85% for HHA-Net), likely reflecting the densely distributed wells and regular groundwater extraction patterns in the NJP that create well-defined spatial associations effectively captured by graph structures. At NJP urban sites, where ST-GNN achieves a lower median MAPE (1.30% vs. 2.18% for HHA-Net), HHA-Net's R^2 quantiles (Q_{25}-Q_{75} = 0.869-0.971) consistently exceed those of ST-GNN (0.845–0.963) across all three quartiles, indicating that HHA-Net provides a consistently more reliable overall fit.

Because NSE is unbounded below, a few extreme outlying sites can severely deflate regional mean NSEs, explaining the sharp discrepancies between low or negative means and vastly superior medians observed across baselines (Knoben et al. 2019, McCuen et al. 2006). For example, CNN-LSTM's mean NSE at YTMR natural sites (-145.23) lies almost two orders of magnitude below its first quartile (-2.10). Similarly, the mean NSEs of ST-GNN (-22.21) and GNN (-11.23) at YTMR agricultural sites are far lower than their medians (-1.08 and -1.28). In the NJP, Transformer's near-zero mean NSE (0.04) is likewise skewed by outlying sites, as

its median NSE reaches 0.80 and its first quartile (0.74) lies well above the mean. Even in terms of medians, baseline NSEs across the YTMR remain negative (-4.57 to -0.14). Furthermore, the interquartile MAPEs of ST-GNN (3.13%-19.86%) and Transformer (1.94%-11.97%) each span more than a sixfold range at YTMR agricultural sites, while Transformer's interquartile MAPE similarly spans 1.89%-27.77% at NJP agricultural sites. In contrast, HHA-Net maintains positive NSEs across its entire interquartile range (Q_{25} = 0.455 and 0.197 at YTMR agricultural and natural sites) and consistently narrow MAPE spans. These comparisons demonstrate that guidance from hydrological domain knowledge significantly enhances model robustness and stability."

"Given the inherent uncertainty in GWD arising from aquifer heterogeneity and spatial variability, we selected an additional 35 sites in the Lower Yangtze River Plain (LYRP), which is a region characterized by intensive human activities, high urbanization levels, and significant industrialization, to validate the model's generalizability and performance (Song et al. 2016). Site information and data are summarized in Table S8, with the corresponding model results presented in Table S9. The results demonstrate that HHA-Net consistently provides high-precision predictions across the three site types, achieving MAPE values ranging from 2.25% to 4.92% and R^2 values ranging from 0.704 to 0.961 (Figure S4). Quantile-based analysis corroborates this consistency, with median (Q50) MAPE values of 3.55%, 3.68%, and 2.22% for agricultural, natural, and urban sites, respectively, and relatively narrow interquartile ranges, further confirming that HHA-Net maintains stable and reliable performance in regions with intensive human disturbance. Overall, HHA-Net achieves consistently excellent performance across diverse hydrogeological scenarios by effectively integrating multi-source heterogeneous data through physics-guided specialized encoders and capturing complex spatiotemporal dependencies via attention mechanisms based on four-dimensional spatial similarity matrices."

Table 2. Performance comparison between HHA-Net and baseline models across agricultural, natural, and urban sites in three study regions.

		Agricultural sites				Natural sites				Urban sites			
		MAPE(%)	RMSE	NSE	R^2	MAPE(%)	RMSE	NSE	R^2	MAPE(%)	RMSE	NSE	R^2
YTMR	GNN	9.808	1.827	-11.230	0.380	10.298	1.100	-8.244	0.253	9.791	1.304	-6.515	0.437
	CNN_LSTM	9.454	1.544	-16.394	0.363	7.390	1.574	-145.233	0.259	7.843	0.999	-0.488	0.255
	MLP	7.780	1.263	-5.034	0.427	6.549	1.029	-24.622	0.285	6.622	1.045	-2.127	0.201
	ST_GNN	10.89	1.781	-22.205	0.242	8.898	1.239	-13.848	0.180	13.967	1.857	-10.805	0.035
	Transformer	8.526	1.010	-1.570	0.454	7.183	0.715	-3.568	0.373	5.628	0.990	-2.431	0.500
	HHA_Net	**3.855**	**0.611**	**0.437**	**0.714**	**2.503**	**0.306**	**0.336**	**0.676**	**3.336**	**0.639**	**0.038**	**0.725**
NCP	GNN	3.979	0.941	0.599	0.856	17.677	1.321	-1.647	0.643	4.640	1.029	0.631	0.760
	CNN_LSTM	2.801	0.676	0.801	0.877	6.316	0.698	0.660	0.754	2.462	0.650	0.892	0.923
	MLP	3.940	0.938	0.736	0.840	12.752	1.189	-0.467	0.710	3.886	0.950	0.647	0.853
	ST_GNN	2.913	0.696	0.797	0.907	**5.004**	0.776	**0.744**	0.813	2.856	0.750	0.848	0.938
	Transformer	1.931	0.496	0.897	0.929	11.304	0.874	-0.194	0.821	1.808	0.488	0.930	0.946
	HHA_Net	**1.097**	**0.296**	**0.966**	**0.978**	***5.947***	**0.459**	***0.691***	**0.943**	**1.027**	**0.302**	**0.975**	**0.984**
NJP	GNN	5.860	0.343	0.606	0.753	2.881	0.403	0.859	0.883	3.819	0.525	0.433	0.736
	CNN_LSTM	7.104	0.374	0.529	0.775	3.573	0.447	0.782	0.862	4.014	0.511	0.476	0.775
	MLP	9.341	0.455	0.141	0.718	3.530	0.475	0.747	0.832	4.473	0.648	0.471	0.712
	ST_GNN	**3.628**	**0.203**	**0.868**	**0.898**	**2.016**	**0.232**	**0.926**	**0.954**	2.601	**0.357**	**0.716**	0.857
	Transformer	16.942	0.713	-1.458	0.774	6.939	0.601	0.040	0.892	6.869	0.730	-0.122	0.738
	HHA_Net	***4.722***	***0.332***	***0.733***	***0.883***	***2.854***	0.432	0.799	***0.923***	**2.525**	***0.394***	***0.525***	**0.892**

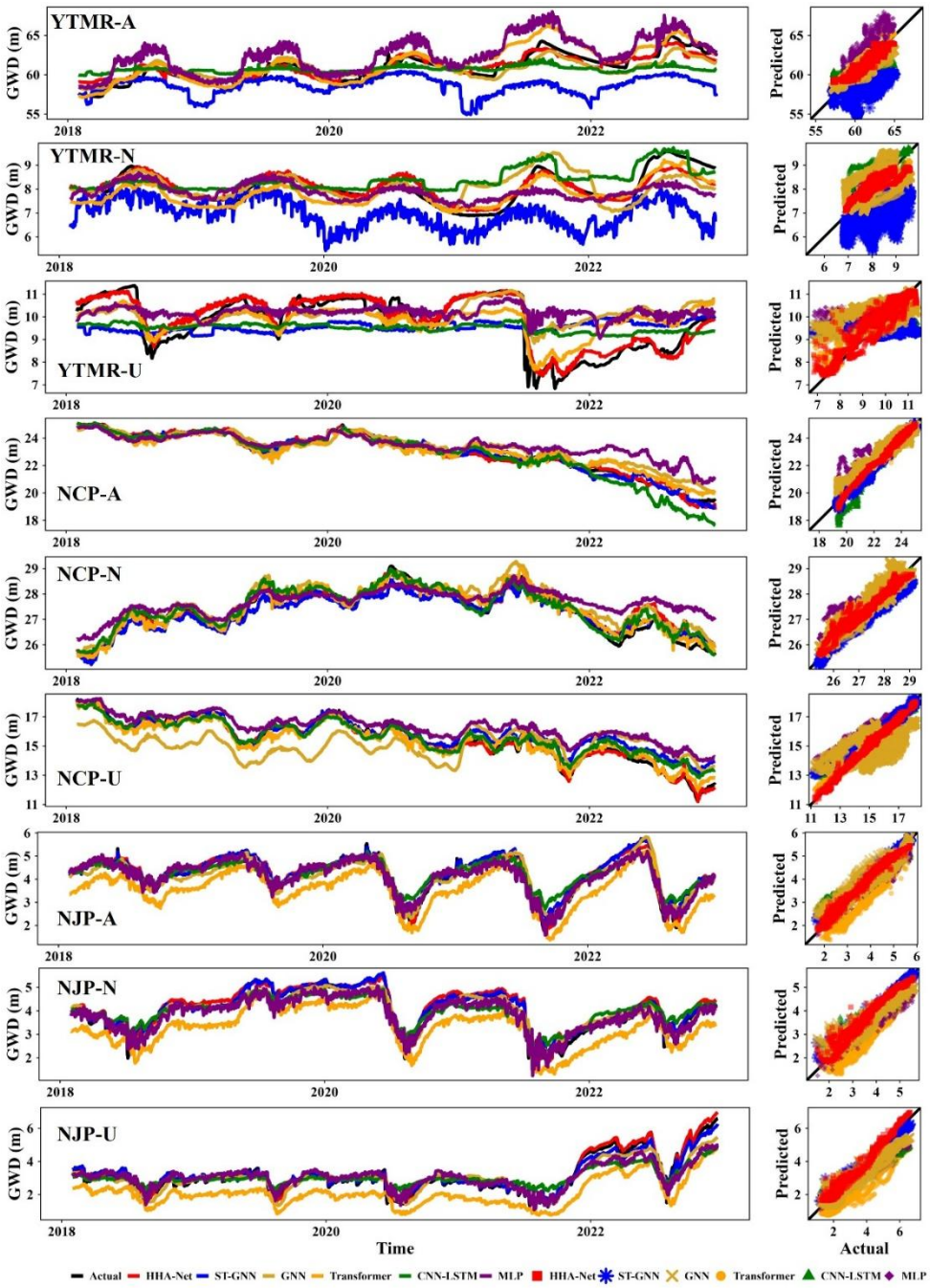

Figure 3: Comparison of GWD prediction performance across the three site types (agricultural (A), natural (N), and urban sites (U)) in the three study regions. Left panels show time series predictions for typical sites in YTMR, NCP, and NJP (prediction performance for all sites is shown in Figure S1, S2 and S3). Right panels display scatter plots comparing predicted versus actual GWD.

Table S5. Optimal hyperparameter configurations in three study regions.

	Model	lr	dropout	Specific Parameters
YTMR	GNN	0.00016	0.238	hidden_size=128
	CNN_LSTM	0.00324	0.102	hidden_size=32
	MLP	0.00041	0.173	hidden_sizes=[128, 64, 32]
	ST_GNN	0.00518	0.217	hidden_size=32
	Transformer	0.00025	0.134	d_model=64, nhead=2,

				num_layers=3
NCP	GNN	0.00958	0.473	hidden_size=32
	CNN_LSTM	0.00010	0.105	hidden_size=32
	MLP	0.00019	0.293	hidden_sizes=[512, 256, 128]
	ST_GNN	0.00060	0.299	hidden_size=64
	Transformer	0.00046	0.349	d_model=128, nhead=8, num_layers=2
NJP	GNN	0.00011	0.102	hidden_size=64
	CNN_LSTM	0.00040	0.173	hidden_size=32
	MLP	0.00015	0.216	hidden_sizes=[512, 256, 128]
	ST_GNN	0.00177	0.245	hidden_size=64
	Transformer	0.00023	0.216	d_model=32, nhead=2, num_layers=2

For the comparison with the GEMS-GER data, it would be helpful to see some more details on how the different encoders were used and how that relates to or differs from the models used on those data sets - that would help to assess whether the performance increase comes from the model architecture or the input data. Also in that comparison, percentiles would be better than averages.

Response

We sincerely appreciate the reviewer for this suggestion. We have revised the GEMS-GER comparison to explicitly clarify that HHA-Net and the CNN and LSTM baselines were provided with an identical set of input variables, and that the human-activity encoder of HHA-Net was disabled due to the lack of socio-economic data, so that HHA-Net used no additional input information beyond what was available to the baselines; and the specific architectural components distinguishing HHA-Net from the baselines, namely its hierarchical, physics-informed encoding and its cross-site spatial attention mechanism, which are absent from the CNN and LSTM models that predict each site independently. This clarification allows us to attribute the observed performance advantage of HHA-Net to its architectural design rather than to additional input data. We have also added percentile-based statistics (median and Q_{25}-Q_{75}) alongside the mean values for both R^2 and NSE, which better reflect model performance given the high spatial heterogeneity across the 3,207 sites in this dataset.

We have revised the sentences in the manuscript as follows:

"Moreover, we further validated the HHA-Net model on the benchmark dataset GEMS-GER, which includes GWD from 3,207 groundwater sites in Germany (Ohmer et al. 2026). To provide a rigorous and fair comparison, HHA-Net and the CNN/LSTM benchmark models were provided with an identical set of input variables, including 7-week antecedent GWD, meteorological forcing (temperature and precipitation), binary extreme-event indicators, and static topographic and site attributes (coordinates, aquifer type, site type, and

watershed). The human-activity encoder of HHA-Net was disabled in this comparison due to the lack of socio-economic data, so that HHA-Net relies only on its historical, hydro-meteorological, and geographical encoders, without using additional input information beyond that available to the baselines. Using 7 weeks of historical GWD and environmental features to predict 1 week ahead, HHA-Net achieved robust predictive performance (mean R^2 = 0.901, median R^2 = 0.954, Q_{25}-Q_{75} = 0.882-0.979), consistently outperforming the CNN baseline (mean R^2 = 0.834, median R^2 = 0.908, Q_{25}-Q_{75} = 0.778-0.958) and LSTM baseline (mean R^2 = 0.877, median R^2 = 0.941, Q_{25}-Q_{75} = 0.849-0.974) (Table S10). Notably, the mean NSEs of the baselines are negative (mean NSE of -2.299 for CNN and -1.806 for LSTM), whereas their median NSE values are 0.763 (Q_{25}-Q_{75} = 0.212-0.915) and 0.826 (Q_{25}-Q_{75} = 0.431-0.938), respectively. In contrast, HHA-Net maintains a positive mean NSE of 0.348 and a superior median NSE of 0.863 (Q_{25}-Q_{75} = 0.633-0.940). This performance gap can be attributed to the architectural design of HHA-Net, which incorporates hierarchical, physics-informed encoding and a cross-site spatial attention mechanism that aggregates information from sites sharing the same prediction date, capabilities that are entirely absent from the CNN and LSTM benchmarks, which predict each site independently based solely on its own historical sequence."

The design decision to predict a 14 week window at once could maybe be motivated a bit better - other approaches predict a single time step in a rolling window approach and for some experiments this is also the case in this study.

Response

We sincerely appreciate the reviewer for this comment, and would like to first clarify that the prediction horizon in this study is 14 days rather than 14 weeks. Regarding the design choice, we agree this deserves clearer motivation and have added an explanation in Section 2.1.6 of the revised manuscript. In brief, the direct multi-step strategy avoids the error accumulation inherent in recursive rolling forecasting over long horizons, allows the composite loss function to jointly optimize the entire predicted sequence through temporal consistency constraints, and provides an immediate and complete medium-range outlook to support timely decision-making. All baseline models were configured with the same direct 14-dimensional output strategy to ensure a fair comparison across all models.

We have added the sentences in the manuscript as follows:

"This study employs a direct multi-step prediction strategy that outputs the complete 14-day sequence in a single forward pass, avoiding the error accumulation inherent in recursive forecasting over long horizons and allowing the composite loss function to jointly optimize the entire predicted trajectory through temporal consistency constraints. A comprehensive medium-range (14 days) prediction provides an immediate and complete outlook on future GWD evolution, supporting groundwater management decisions."

The idea to combine multiple loss functions is interesting and could be investigated in a bit more detail. Did the authors observe that the results changed significantly without that multi-objective optimization, or with different weights of the individual loss functions? How were the weights optimized? Do the relative differences between models change with the weightings and without the multiple losses?

Response

We sincerely thank the reviewer for this constructive and insightful comment. To address this concern, we have conducted dedicated sensitivity analysis to systematically evaluate the impact of the multi-objective loss formulation and different weight configurations. Results reveal that optimizing solely with mean squared error leads to severe performance degradation, whereas introducing temporal and physical consistency constraints markedly enhances stability and model accuracy. To optimize the weighting scheme, we empirically set the initial loss weights to balance primary data fitting against outlier robustness and implemented a progressive scheduling strategy that gradually strengthens the consistency constraints as training proceeds, which achieved the best overall balance across all metrics compared with any static weighting combination. Crucially, the relative performance advantage of HHA-Net over all baseline models remains consistently solid across every tested loss configuration, and even the single-loss configuration of HHA-Net outperforms the best baseline, demonstrating that the primary predictive superiority originates from the architectural design while the multi-objective formulation and progressive scheduling further refine and stabilize this advantage.

We have added the corresponding sentences in the manuscript as follows:

"The total loss function is formulated as a weighted linear combination of mean squared error (MSE) loss, mean absolute error (MAE) loss, Huber loss, and a consistency constraint loss with distinct weight coefficients, forming a multi-level adaptive optimization objective (Appendix A4). The sensitivity of model performance to the loss weights and to the progressive schedule is systematically evaluated through dedicated experiments. The architectural hyperparameters of HHA-Net, including a hidden dimension of 64, four attention heads, multi-scale convolutional kernel sizes of 3, 5, and 7 days, and the funnel-shaped prediction network layers, were selected to balance the representation capacity across multi-source features with model generalizability while preventing overfitting. Key physical parameters such as maximum response delay, spatial similarity initializations, and site-type bias distributions were directly guided by hydrogeological domain knowledge and prior empirical findings."

"To further address the sensitivity of model performance to the design of the multi-objective loss function, model performance was compared under different loss weight configurations while keeping the model architecture and all other training settings identical. The tested configurations included the MSE term alone, removing the consistency constraints entirely (Consistency=0), fixing the consistency weight at 0.10, 0.15, or 0.20 without progressive scheduling, an MAE-dominant weighting scheme, and static weights applied throughout training (Table

S6 and S7). Across all regions and site types, the MSE-only configuration consistently performs the worst, with markedly higher MAPE and, in several cases, negative mean NSE (e.g., -0.889 at YTMR agricultural sites). Introducing the consistency constraint, even without scheduling, substantially improves NSE and R^2 relative to the MSE-only case (e.g., mean NSE rising from -0.889 to 0.157-0.449 at YTMR agricultural sites). Among the fixed configurations, no single weighting scheme is uniformly optimal, while the progressive weighting schedule adopted in HHA-Net, achieves the best overall balance across metrics (MAPE = 3.096%, RMSE = 0.419, NSE = 0.611, R^2 = 0.857), indicating that gradually strengthening the consistency constraint during training yields more robust generalization. Importantly, every tested loss weight configuration of HHA-Net still substantially outperforms all baseline models; even MSE-only HHA-Net (MAPE = 3.996%, NSE = 0.405) surpasses the best baseline, demonstrating that the performance advantage of HHA-Net stems primarily from its architectural design rather than loss-weight tuning, while the multi-objective formulation and progressive scheduling further refine and stabilize this advantage."

"These three error terms are applied to the standardized 14-day prediction sequence, with each prediction step weighted by an exponential temporal-decay coefficient. The temporal consistency loss includes first- and second-order difference losses: the former constrains prediction sequence continuity by calculating differences between adjacent time-step predictions, whereas the latter constrains smoothness by calculating curvature changes in the prediction sequence, thereby ensuring that prediction results conform to physical process continuity requirements and avoiding physically unreasonable predictions from purely data-driven methods.

The physical consistency loss primarily includes variation magnitude constraints and continuity constraints. Variation magnitude constraints penalize excessively large day-to-day fluctuations in the predicted sequence, including both single-step and cumulative two-step changes. Continuity constraints ensure reasonable continuity between the initial predicted values and the latest historical observations, preventing unrealistic abrupt changes. In the total loss function, the initial weights are determined empirically. The primary MSE loss is assigned the largest weight (0.50) as the main optimization objective; the MAE (0.30) and Huber (0.15) losses provide balancing and stabilizing effects against outlier sites; and the temporal and physical consistency constraints share a joint weight (0.05) to ensure the temporal and physical plausibility of predictions without excessively constraining model flexibility. During training, these weights follow a progressive scheduling strategy updated every 20 epochs. In the early training stage, the consistency weight remains small so that the model focuses on fitting the observed data; in the middle and later stages, the consistency weight is gradually increased to 0.10 and then to 0.20, while the MSE weight is moderately reduced, thereby progressively strengthening physical plausibility as the model converges."

Table S6. Performance comparison between HHA-Net and different weights of loss functions across agricultural, natural, and urban sites in three study regions based on mean and quantiles (25%/50%/75%) metrics.

Area		MAPE(%)				RMSE				NSE				R^2			
	Agricultural	Mean	Q_{25}	Q_{50}	Q_{75}	Mean	Q_{25}	Q_{50}	Q_{75}	Mean	Q_{25}	Q_{50}	Q_{75}	Mean	Q_{25}	Q_{50}	Q_{75}
YTMR	MSE only	7.823	1.445	3.191	11.318	0.813	0.591	0.833	0.990	-0.889	-1.340	0.233	0.635	0.710	0.638	0.709	0.816
	Consistency=0	4.479	1.501	3.050	7.518	0.681	0.375	0.583	0.908	0.271	0.210	0.446	0.561	0.686	0.596	0.716	0.849
	Consistency=0.10	4.132	1.559	2.819	6.229	0.686	0.382	0.645	0.952	0.327	0.051	0.476	0.594	0.744	0.703	0.751	0.814
	Consistency=0.15	5.452	1.691	3.977	6.664	0.786	0.498	0.694	1.100	0.157	-0.005	0.262	0.551	0.715	0.647	0.730	0.765
	Consistency=0.20	4.590	1.184	2.663	8.261	0.634	0.453	0.658	0.807	0.201	0.157	0.436	0.703	0.688	0.596	0.657	0.802
	MAE-dominant	4.424	1.082	2.878	7.436	0.609	0.361	0.454	0.806	0.449	0.397	0.544	0.709	0.744	0.715	0.762	0.802
	Static weights	7.280	1.312	2.706	13.059	0.721	0.550	0.661	0.826	-0.325	-0.458	0.465	0.693	0.738	0.644	0.755	0.825
	Natural																
	MSE only	4.057	1.275	1.664	5.723	0.372	0.258	0.318	0.464	0.096	-0.503	0.321	0.549	0.619	0.515	0.680	0.787
	Consistency=0	3.384	1.860	2.685	3.769	0.362	0.296	0.346	0.450	0.102	0.106	0.403	0.581	0.616	0.524	0.663	0.776
	Consistency=0.10	2.813	1.038	2.084	3.685	0.314	0.222	0.338	0.364	0.388	0.285	0.538	0.681	0.670	0.578	0.740	0.805
	Consistency=0.15	2.956	1.663	2.746	3.416	0.366	0.184	0.308	0.352	-0.024	0.213	0.508	0.675	0.655	0.610	0.717	0.801
	Consistency=0.20	4.085	1.052	2.067	6.456	0.363	0.244	0.347	0.482	0.201	-0.048	0.458	0.571	0.623	0.561	0.664	0.772
	MAE-dominant	3.318	0.914	1.840	4.834	0.299	0.220	0.271	0.399	0.321	0.061	0.507	0.782	0.680	0.618	0.759	0.846
	Static weights	4.184	1.673	2.937	6.530	0.422	0.374	0.435	0.485	-0.184	-0.515	-0.103	0.391	0.632	0.594	0.665	0.801
	Urban																
	MSE only	5.041	2.627	4.249	5.197	0.810	0.505	0.530	0.606	-0.543	-0.121	0.494	0.548	0.727	0.670	0.761	0.858
	Consistency=0	4.900	2.890	3.937	5.774	0.753	0.470	0.554	0.581	0.329	0.131	0.363	0.649	0.589	0.588	0.626	0.672
	Consistency=0.10	4.181	2.134	3.408	4.523	0.716	0.421	0.454	0.489	0.076	0.050	0.519	0.747	0.633	0.495	0.686	0.832
	Consistency=0.15	3.766	2.072	4.015	4.393	0.709	0.391	0.432	0.471	-0.024	0.063	0.648	0.772	0.745	0.687	0.789	0.894
	Consistency=0.20	5.699	2.545	3.873	5.106	0.840	0.446	0.531	0.757	0.152	-0.287	0.387	0.641	0.584	0.468	0.625	0.775

	MAE-dominant	3.258	1.609	2.646	3.379	0.563	0.273	0.382	0.397	0.582	0.393	0.684	0.836	0.658	0.507	0.778	0.919
	Static weights	3.839	2.060	4.334	4.994	0.662	0.392	0.477	0.533	0.484	0.285	0.613	0.754	0.620	0.451	0.680	0.828
	Agricultural																
	MSE only	1.713	1.187	1.431	2.107	0.455	0.328	0.390	0.519	0.912	0.885	0.951	0.986	0.962	0.940	0.978	0.992
	Consistency=0	1.548	0.903	1.174	1.793	0.397	0.295	0.373	0.479	0.930	0.901	0.959	0.987	0.959	0.927	0.979	0.991
	Consistency=0.10	1.246	0.754	1.023	1.538	0.338	0.238	0.278	0.417	0.947	0.932	0.962	0.991	0.977	0.963	0.983	0.995
	Consistency=0.15	1.111	0.703	0.933	1.309	0.289	0.194	0.267	0.359	0.967	0.941	0.971	0.994	0.979	0.963	0.987	0.995
	Consistency=0.20	1.012	0.616	0.719	1.024	0.271	0.185	0.219	0.344	0.969	0.951	0.981	0.995	0.980	0.966	0.986	0.997
	MAE-dominant	1.439	0.851	1.110	1.384	0.368	0.240	0.342	0.479	0.944	0.920	0.969	0.985	0.971	0.955	0.982	0.992
	Static weights	1.007	0.626	0.780	1.115	0.264	0.201	0.233	0.299	0.969	0.950	0.980	0.992	0.978	0.967	0.989	0.996
	Natural																
	MSE only	6.832	1.421	1.711	2.649	0.586	0.461	0.663	0.738	0.614	0.748	0.942	0.973	0.901	0.807	0.975	0.983
	Consistency=0	5.193	1.151	1.174	2.107	0.483	0.328	0.572	0.575	0.737	0.812	0.919	0.983	0.929	0.852	0.974	0.987
	Consistency=0.10	5.542	0.933	1.162	1.701	0.437	0.335	0.462	0.482	0.747	0.877	0.949	0.988	0.797	0.927	0.982	0.991
NCP	Consistency=0.15	5.923	0.922	1.266	1.506	0.421	0.359	0.390	0.439	0.716	0.912	0.974	0.990	0.964	0.940	0.987	0.991
	Consistency=0.20	5.715	0.747	0.779	1.722	0.391	0.232	0.357	0.437	0.733	0.927	0.957	0.990	0.905	0.947	0.988	0.993
	MAE-dominant	5.758	1.098	1.373	1.817	0.461	0.298	0.474	0.579	0.722	0.808	0.930	0.989	0.836	0.898	0.971	0.989
	Static weights	5.429	0.685	0.780	1.297	0.376	0.241	0.318	0.383	0.753	0.936	0.942	0.992	0.859	0.958	0.98	0.994
	Urban																
	MSE only	1.837	1.196	1.421	1.896	0.458	0.272	0.500	0.598	0.935	0.906	0.959	0.966	0.969	0.962	0.976	0.984
	Consistency=0	1.483	1.035	1.273	1.462	0.405	0.314	0.402	0.479	0.947	0.927	0.964	0.974	0.966	0.968	0.977	0.983
	Consistency=0.10	1.136	0.782	1.166	1.335	0.315	0.216	0.264	0.389	0.973	0.965	0.970	0.982	0.988	0.986	0.989	0.991
	Consistency=0.15	0.906	0.680	0.887	1.017	0.244	0.200	0.241	0.303	0.983	0.976	0.985	0.987	0.989	0.988	0.991	0.991
	Consistency=0.20	0.772	0.616	0.727	0.814	0.222	0.146	0.224	0.279	0.983	0.981	0.987	0.992	0.992	0.990	0.993	0.995
	MAE-dominant	1.400	0.932	1.307	1.589	0.390	0.249	0.326	0.486	0.956	0.939	0.955	0.979	0.979	0.975	0.981	0.989
	Static weights	0.891	0.633	0.759	0.888	0.230	0.168	0.231	0.303	0.983	0.978	0.985	0.990	0.991	0.991	0.992	0.993

NJP	*Agricultural*																
	MSE only	3.914	1.068	2.046	5.370	0.239	0.163	0.204	0.273	0.839	0.734	0.904	0.967	0.916	0.859	0.945	0.975
	Consistency=0	5.313	1.338	3.703	6.633	0.291	0.242	0.259	0.337	0.693	0.554	0.802	0.954	0.898	0.870	0.926	0.979
	Consistency=0.10	4.661	1.129	2.060	6.789	0.273	0.194	0.231	0.323	0.802	0.672	0.880	0.955	0.909	0.851	0.929	0.974
	Consistency=0.15	3.961	1.282	2.430	5.006	0.260	0.182	0.225	0.291	0.804	0.697	0.871	0.965	0.893	0.841	0.946	0.974
	Consistency=0.20	3.119	1.166	2.085	3.868	0.214	0.146	0.199	0.260	0.872	0.822	0.904	0.969	0.943	0.912	0.958	0.986
	MAE-dominant	4.439	0.777	1.892	5.628	0.237	0.179	0.220	0.245	0.820	0.748	0.860	0.953	0.927	0.877	0.941	0.978
	Static weights	4.782	0.960	2.577	4.717	0.255	0.166	0.249	0.329	0.655	0.782	0.915	0.977	0.910	0.880	0.962	0.984
	Natural																
	MSE only	2.484	1.098	1.532	2.225	0.353	0.187	0.261	0.526	0.852	0.820	0.936	0.959	0.938	0.937	0.973	0.979
	Consistency=0	2.652	0.925	1.111	1.994	0.296	0.229	0.277	0.347	0.870	0.886	0.942	0.975	0.935	0.945	0.972	0.983
	Consistency=0.10	2.597	1.080	1.634	2.327	0.348	0.232	0.302	0.430	0.844	0.841	0.926	0.964	0.915	0.907	0.966	0.978
	Consistency=0.15	2.355	1.195	1.592	2.486	0.354	0.231	0.264	0.409	0.846	0.804	0.949	0.964	0.922	0.832	0.971	0.980
	Consistency=0.20	1.797	0.846	1.367	1.890	0.302	0.185	0.252	0.372	0.912	0.892	0.954	0.968	0.950	0.924	0.978	0.987
	MAE-dominant	2.286	0.740	1.090	2.147	0.304	0.198	0.237	0.317	0.889	0.899	0.955	0.973	0.937	0.953	0.970	0.982
	Static weights	2.383	0.991	1.294	1.887	0.305	0.210	0.277	0.362	0.838	0.880	0.949	0.972	0.938	0.923	0.967	0.986
	Urban																
	MSE only	2.266	0.680	1.424	3.191	0.312	0.185	0.226	0.370	0.834	0.827	0.900	0.963	0.883	0.909	0.958	0.973
	Consistency=0	2.318	0.902	1.423	2.770	0.303	0.226	0.292	0.369	0.787	0.780	0.913	0.973	0.886	0.901	0.960	0.979
	Consistency=0.10	2.805	0.902	1.887	3.057	0.340	0.217	0.297	0.459	0.683	0.744	0.885	0.971	0.848	0.866	0.946	0.974
	Consistency=0.15	2.559	1.222	1.978	3.210	0.355	0.247	0.318	0.468	0.506	0.811	0.880	0.945	0.874	0.874	0.945	0.969
	Consistency=0.20	1.980	0.699	1.380	2.67	0.288	0.167	0.237	0.400	0.820	0.848	0.915	0.974	0.904	0.906	0.957	0.981
	MAE-dominant	2.108	0.713	1.266	2.819	0.286	0.175	0.237	0.358	0.840	0.837	0.910	0.974	0.909	0.905	0.947	0.980
	Static weights	2.379	0.845	1.561	2.220	0.294	0.213	0.314	0.366	0.791	0.830	0.921	0.951	0.881	0.914	0.948	0.977

Table S7. Summary of performance comparison between HHA-Net (different weights of loss functions) and baseline models.

	MAPE(%)	RMSE	NSE	R^2
HHA_Net	3.096	0.419	0.611	0.857
MSE only	3.996	0.488	0.405	0.847
Consistency=0	3.474	0.441	0.629	0.829
Consistency=0.10	3.234	0.418	0.643	0.831
Consistency=0.15	3.221	0.420	0.547	0.859
Consistency=0.20	3.196	0.391	0.649	0.841
MAE-dominant	3.158	0.390	0.724	0.849
Static weights	3.574	0.392	0.551	0.838
GNN	7.639	0.977	-2.723	0.633
CNN_LSTM	5.661	0.830	-17.552	0.649
MLP	6.541	0.888	-3.278	0.619
ST_GNN	5.863	0.876	-4.662	0.647
Transformer	7.458	0.735	-0.830	0.714

The large differences between the proposed model and all baselines seem at odds with existing literature and prior empirical evidence from other geographical regions. Also the very low NSE values for some geographical regions appear puzzling, just like the differences between NSE and R^2 (which are often used synonymously and which often are considered equivalent under some assumptions - it's good that the authors comment on this). With the presentation of quantiles and HPO this should change and will improve the extent to which one can assess the results and their validity.

Response

We sincerely appreciate the reviewer for this insightful comment. To address it, we have performed an independent Optuna-based hyperparameter optimization for each baseline model in each region. After optimization, most baselines now achieve positive mean NSE values in the NCP and NJP, for example 0.930 for the Transformer at NCP urban sites and 0.926 for ST-GNN at NJP natural sites, whereas the remaining low or negative mean NSE values are concentrated in the YTMR. To clarify the remaining low or negative NSE values and their discrepancies with R^2, we added comprehensive quantile metrics (Q_{25}, Q_{50}, and Q_{75}) for all models and regions. Because NSE is bounded above at 1 but unbounded below, a few extreme outlying sites can severely deflate regional arithmetic means while median values and R^2 remain considerably higher. For instance, the mean NSE of CNN-LSTM at YTMR natural sites (-145.23) lies almost two orders of magnitude below its own first quartile (-2.10), despite a median NSE of -0.89, and similarly the Transformer in the NJP achieves a median NSE of 0.80 despite a deflated mean of 0.04. Furthermore, we conducted sensitivity experiments across various multi-objective loss configurations, confirming that the performance gains of HHA-Net stem fundamentally from its domain-guided architectural design rather than loss-weight tuning. We believe that the combination of rigorous hyperparameter

optimization, quantile-based reporting, and structural sensitivity analysis now allows a transparent assessment of the results and their validity, exactly as the reviewer suggests. We have revised the section 4.1.1 (HHA-Net performance evaluation and benchmark comparison) in the manuscript as follows:

"HHA-Net exhibits excellent predictive capability across various hydrological conditions (Figure 3 and Figure S1-S3), effectively capturing complex seasonal oscillations, gradual/abrupt change trends, and extreme values. Scatter plots demonstrate that HHA-Net achieves better alignment along the 1:1 diagonal line with tighter clustering, while baseline models exhibit greater scatter and larger prediction errors across different GWD ranges. HHA-Net exhibits different performance across site types (Table 2). At urban sites, HHA-Net achieves the MAPE of 1.03%-3.34% and RMSE of 0.30-0.64, reaching its optimal performance in the NCP (R^2 = 0.984, NSE = 0.975). For natural sites, the model maintains MAPE within 2.50%-5.95%, RMSE between 0.31 and 0.46, and R^2 ranging from 0.676 to 0.943. In the YTMR, the interquartile MAPE range (1.29%-3.59%) closely brackets the mean, indicating relatively uniform performance across the region; whereas for natural sites in NCP the mean MAPE of 5.95% considerably exceeds both the median (1.18%) and the third quartile (1.55%), demonstrating that the elevated mean is driven by some sites with anomalously large errors. Agricultural sites maintain competitive prediction accuracy, with MAPE ranging from 1.10% to 4.72%. HHA-Net reaches its highest agricultural MAPE (4.72%) in the NJP region. This specific challenge may be attributed to multiple factors including rainfall infiltration, irrigation withdrawal, and soil permeability changes induced by agricultural activities, resulting in high spatiotemporal heterogeneity in groundwater systems at agricultural sites (Zhang et al. 2019).

The optimal hyperparameter configurations and predictive performance for baseline models, obtained through the Optuna-based hyperparameter optimization procedure, are summarized in Table S4 and S5. With these configurations, most baselines achieve positive mean NSE values in the NCP and NJP (e.g., 0.930 for Transformer at NCP urban sites and 0.926 for ST-GNN at NJP natural sites), whereas low or negative mean NSE values are concentrated in the YTMR. Overall, HHA-Net outperforms baselines across most regions and site types, particularly in the YTMR and NCP. At agricultural sites, HHA-Net reduces prediction errors compared to the top baselines (e.g., MAPE of 1.10% vs. Transformer's 1.93% in NCP; 3.86% vs. MLP's 7.78% in YTMR). At YTMR natural sites, GNN (10.30%) and Transformer (7.18%) exhibit mean MAPEs roughly 3 to 4 times higher than those of HHA-Net (2.50%), with wider and more right-skewed quantile ranges (e.g., 2.54%-17.51% for GNN) compared to those of HHA-Net (1.29%-3.59%). At YTMR urban sites, HHA-Net similarly yields major improvements, with ST-GNN's MAPE (13.97%) exceeding four times that of HHA-Net (3.34%). An exception occurs in the NJP region, where ST-GNN demonstrates superior mean performance at agricultural and natural sites (MAPE = 3.63% and 2.02%, vs. 4.72% and 2.85% for HHA-Net), likely reflecting the densely distributed wells and regular groundwater extraction patterns in the

NJP that create well-defined spatial associations effectively captured by graph structures. At NJP urban sites, where ST-GNN achieves a lower median MAPE (1.30% vs. 2.18% for HHA-Net), HHA-Net's R^2 quantiles (Q_{25}-Q_{75} = 0.869-0.971) consistently exceed those of ST-GNN (0.845–0.963) across all three quartiles, indicating that HHA-Net provides a consistently more reliable overall fit.

Because NSE is unbounded below, a few extreme outlying sites can severely deflate regional mean NSEs, explaining the sharp discrepancies between low or negative means and vastly superior medians observed across baselines (Knoben et al. 2019, McCuen et al. 2006). For example, CNN-LSTM's mean NSE at YTMR natural sites (-145.23) lies almost two orders of magnitude below its first quartile (-2.10). Similarly, the mean NSEs of ST-GNN (-22.21) and GNN (-11.23) at YTMR agricultural sites are far lower than their medians (-1.08 and -1.28). In the NJP, Transformer's near-zero mean NSE (0.04) is likewise skewed by outlying sites, as its median NSE reaches 0.80 and its first quartile (0.74) lies well above the mean. Even in terms of medians, baseline NSEs across the YTMR remain negative (-4.57 to -0.14). Furthermore, the interquartile MAPEs of ST-GNN (3.13%-19.86%) and Transformer (1.94%-11.97%) each span more than a sixfold range at YTMR agricultural sites, while Transformer's interquartile MAPE similarly spans 1.89%-27.77% at NJP agricultural sites. In contrast, HHA-Net maintains positive NSEs across its entire interquartile range (Q_{25} = 0.455 and 0.197 at YTMR agricultural and natural sites) and consistently narrow MAPE spans. These comparisons demonstrate that guidance from hydrological domain knowledge significantly enhances model robustness and stability.

To further address the sensitivity of model performance to the design of the multi-objective loss function, model performance was compared under different loss weight configurations while keeping the model architecture and all other training settings identical. The tested configurations included the MSE term alone, removing the consistency constraints entirely (Consistency=0), fixing the consistency weight at 0.10, 0.15, or 0.20 without progressive scheduling, an MAE-dominant weighting scheme, and static weights applied throughout training (Table S6 and S7). Across all regions and site types, the MSE-only configuration consistently performs the worst, with markedly higher MAPE and, in several cases, negative mean NSE (e.g., -0.889 at YTMR agricultural sites). Introducing the consistency constraint, even without scheduling, substantially improves NSE and R^2 relative to the MSE-only case (e.g., mean NSE rising from -0.889 to 0.157-0.449 at YTMR agricultural sites). Among the fixed configurations, no single weighting scheme is uniformly optimal, while the progressive weighting schedule adopted in HHA-Net, achieves the best overall balance across metrics (MAPE = 3.096%, RMSE = 0.419, NSE = 0.611, R^2 = 0.857), indicating that gradually strengthening the consistency constraint during training yields more robust generalization. Importantly, every tested loss weight configuration of HHA-Net still substantially outperforms all baseline models; even MSE-only HHA-Net (MAPE = 3.996%, NSE = 0.405) surpasses the best baseline, demonstrating that the performance advantage of HHA-Net stems primarily from its architectural design rather than

loss-weight tuning, while the multi-objective formulation and progressive scheduling further refine and stabilize this advantage.

Given the inherent uncertainty in GWD arising from aquifer heterogeneity and spatial variability, we selected an additional 35 sites in the Lower Yangtze River Plain (LYRP), which is a region characterized by intensive human activities, high urbanization levels, and significant industrialization, to validate the model's generalizability and performance (Song et al. 2016). Site information and data are summarized in Table S8, with the corresponding model results presented in Table S9. The results demonstrate that HHA-Net consistently provides high-precision predictions across the three site types, achieving MAPE values ranging from 2.25% to 4.92% and R^2 values ranging from 0.704 to 0.961 (Figure S4). Quantile-based analysis corroborates this consistency, with median MAPE values of 3.55%, 3.68%, and 2.22% for agricultural, natural, and urban sites, respectively, and relatively narrow interquartile ranges, further confirming that HHA-Net maintains stable and reliable performance in regions with intensive human disturbance. Overall, HHA-Net achieves consistently excellent performance across diverse hydrogeological scenarios by effectively integrating multi-source heterogeneous data through physics-guided specialized encoders and capturing complex spatiotemporal dependencies via attention mechanisms based on four-dimensional spatial similarity matrices.

Moreover, we further validated the HHA-Net model on the benchmark dataset GEMS-GER, which includes GWD from 3,207 groundwater sites in Germany (Ohmer et al. 2026). To provide a rigorous and fair comparison, HHA-Net and the CNN/LSTM benchmark models were provided with an identical set of input variables, including 7-week antecedent GWD, meteorological forcing (temperature and precipitation), binary extreme-event indicators, and static topographic and site attributes (coordinates, aquifer type, site type, and watershed). The human-activity encoder of HHA-Net was disabled in this comparison due to the lack of socio-economic data, so that HHA-Net relies only on its historical, hydro-meteorological, and geographical encoders, without using additional input information beyond that available to the baselines. Using 7 weeks of historical GWD and environmental features to predict 1 week ahead, HHA-Net achieved robust predictive performance (mean $R^2 = 0.901$, median $R^2 = 0.954$, Q_{25}-$Q_{75} = 0.882$-0.979), consistently outperforming the CNN baseline (mean $R^2 = 0.834$, median $R^2 = 0.908$, Q_{25}-$Q_{75} = 0.778$-0.958) and LSTM baseline (mean $R^2 = 0.877$, median $R^2 = 0.941$, Q_{25}-$Q_{75} = 0.849$-0.974) (Table S10). Notably, the mean NSEs of the baselines are negative (mean NSE of -2.299 for CNN and -1.806 for LSTM), whereas their median NSE values are 0.763 (Q_{25}-$Q_{75} = 0.212$-0.915) and 0.826 (Q_{25}-$Q_{75} = 0.431$-0.938), respectively. In contrast, HHA-Net maintains a positive mean NSE of 0.348 and a superior median NSE of 0.863 (Q_{25}-$Q_{75} = 0.633$-0.940). This performance gap can be attributed to the architectural design of HHA-Net, which incorporates hierarchical, physics-informed encoding and a cross-site spatial attention mechanism that aggregates information from sites sharing the same prediction date, capabilities that are entirely absent from the CNN and LSTM

benchmarks, which predict each site independently based solely on its own historical sequence. The distributions of model metrics and examples of scatter plots comparing predicted and observed GWD are shown in the Figure S5."

Table 2. Performance comparison between HHA-Net and baseline models across agricultural, natural, and urban sites in three study regions.

		Agricultural sites				Natural sites				Urban sites			
		MAPE(%)	RMSE	NSE	R^2	MAPE(%)	RMSE	NSE	R^2	MAPE(%)	RMSE	NSE	R^2
YTMR	GNN	9.808	1.827	-11.230	0.380	10.298	1.100	-8.244	0.253	9.791	1.304	-6.515	0.437
	CNN_LSTM	9.454	1.544	-16.394	0.363	7.390	1.574	-145.233	0.259	7.843	0.999	-0.488	0.255
	MLP	7.780	1.263	-5.034	0.427	6.549	1.029	-24.622	0.285	6.622	1.045	-2.127	0.201
	ST_GNN	10.89	1.781	-22.205	0.242	8.898	1.239	-13.848	0.180	13.967	1.857	-10.805	0.035
	Transformer	8.526	1.010	-1.570	0.454	7.183	0.715	-3.568	0.373	5.628	0.990	-2.431	0.500
	HHA_Net	**3.855**	**0.611**	**0.437**	**0.714**	**2.503**	**0.306**	**0.336**	**0.676**	**3.336**	**0.639**	**0.038**	**0.725**
NCP	GNN	3.979	0.941	0.599	0.856	17.677	1.321	-1.647	0.643	4.640	1.029	0.631	0.760
	CNN_LSTM	2.801	0.676	0.801	0.877	6.316	0.698	0.660	0.754	2.462	0.650	0.892	0.923
	MLP	3.940	0.938	0.736	0.840	12.752	1.189	-0.467	0.710	3.886	0.950	0.647	0.853
	ST_GNN	2.913	0.696	0.797	0.907	**5.004**	0.776	**0.744**	0.813	2.856	0.750	0.848	0.938
	Transformer	1.931	0.496	0.897	0.929	11.304	0.874	-0.194	0.821	1.808	0.488	0.930	0.946
	HHA_Net	**1.097**	**0.296**	**0.966**	**0.978**	***5.947***	**0.459**	***0.691***	**0.943**	**1.027**	**0.302**	**0.975**	**0.984**
NJP	GNN	5.860	0.343	0.606	0.753	2.881	0.403	0.859	0.883	3.819	0.525	0.433	0.736
	CNN_LSTM	7.104	0.374	0.529	0.775	3.573	0.447	0.782	0.862	4.014	0.511	0.476	0.775
	MLP	9.341	0.455	0.141	0.718	3.530	0.475	0.747	0.832	4.473	0.648	0.471	0.712
	ST_GNN	**3.628**	**0.203**	**0.868**	**0.898**	**2.016**	**0.232**	**0.926**	**0.954**	2.601	**0.357**	**0.716**	0.857
	Transformer	16.942	0.713	-1.458	0.774	6.939	0.601	0.040	0.892	6.869	0.730	-0.122	0.738
	HHA_Net	***4.722***	***0.332***	***0.733***	***0.883***	***2.854***	0.432	0.799	***0.923***	**2.525**	***0.394***	***0.525***	**0.892**

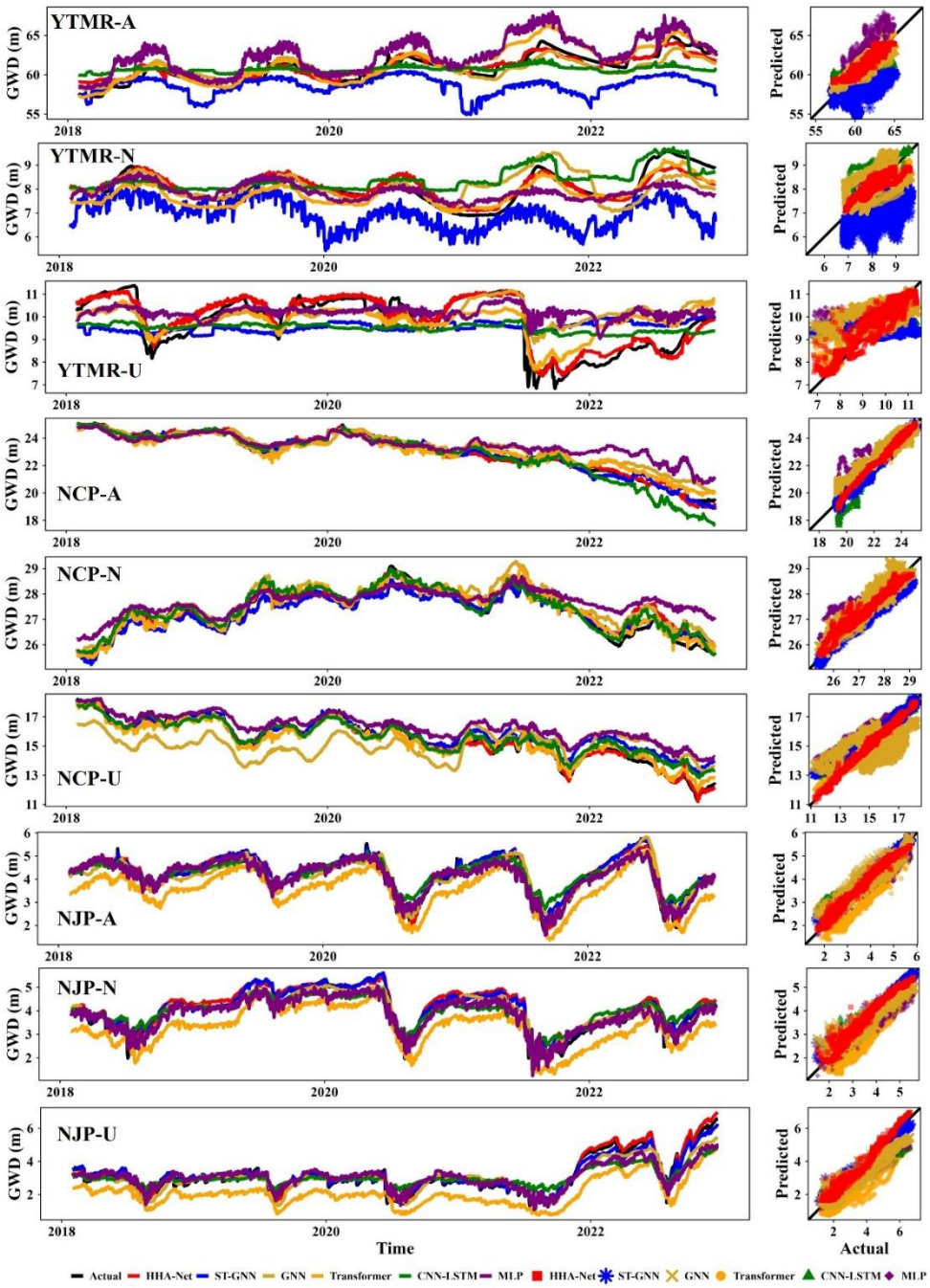

Figure 3: Comparison of GWD prediction performance across the three site types (agricultural (A), natural (N), and urban sites (U)) in the three study regions. Left panels show time series predictions for typical sites in YTMR, NCP, and NJP (prediction performance for all sites is shown in Figure S1, S2 and S3). Right panels display scatter plots comparing predicted versus actual GWD.

Minor remarks:
Line 268: „Euclidean data" - maybe this term could be explained a bit more, I guess neural networks and in fact all of the baseline methods are able to model data on nonlinear manifolds, too, if that's what the authors refer to with Euclidean data?

Response

We sincerely appreciate the reviewer for the comment. We would like to clarify that Euclidean data here refers to the structural topology of the input data (e.g., fixed grids or sequences), rather than the model's capacity to learn nonlinear relationships. Traditional architectures such as CNNs and LSTMs operate on data with fixed, regular structures, whereas GNNs are designed to directly handle graph-structured data with arbitrary, irregular topological connectivity defined by explicit edges.

We have revised the corresponding sentences in the manuscript as follows:

"GNNs are designed to overcome the limitations of traditional neural networks, which are primarily built for data with fixed, regular structures (e.g., grids or sequences) and thus cannot directly operate on graph-structured data with irregular topological connectivity. Their core mechanism enables each node to update its representation by aggregating feature information from neighboring nodes as defined by the graph structure (Asif et al. 2021, Bai and Tahmasebi 2023)."

How were the architectural parameters chosen, was there some neural architecture search, hyper parameter optimization or were the choices empirically motivated, or by some prior work?

Response

We sincerely appreciate the reviewer for this insightful comment. The architectural parameters of HHA-Net were determined through a combination of hydrogeological domain knowledge, prior literature, and empirical design. Specifically, physical constraints such as the maximum lag time, spatial similarity initializations, and site type prior weights were directly formulated based on established hydrological principles and aquifer response characteristics. Meanwhile, structural dimensions including the latent hidden dimension of sixty four, four attention heads, multi scale convolutional kernels of three, five, and seven days, and the progressive prediction layers were empirically configured to balance multi source feature representation with model generalizability. The rationale and necessity of these architectural configurations were systematically confirmed through the hierarchical ablation experiments across all study regions. We have added the sentences in the manuscript as follows:

"The architectural hyperparameters of HHA-Net, including a hidden dimension of 64, four attention heads, multi-scale convolutional kernel sizes of 3, 5, and 7 days, and the funnel-shaped prediction network layers, were selected to balance the representation capacity across multi-source features with model generalizability while preventing overfitting. Key physical parameters such as maximum response delay, spatial similarity initializations, and site-type bias distributions were directly guided by hydrogeological domain knowledge and prior empirical findings."

"Encoder-level ablation reveals substantial performance degradation when removing individual encoders, with the historical encoder demonstrating the most critical role across all regions (Table S11). Removing the historical encoder results in RMSE increases of 2213.2%, 1490.2%, and 1398.5% in the YTMR, NCP, and

NJP, respectively, indicating that historical GWD states serve as the foundation for accurate predictions. This finding is further corroborated by the quantile-based RMSE, with the median RMSE increasing by more than an order of magnitude and the interquartile range (Q_{25}–Q_{75}) also shifting substantially upward across all three regions. Removal of the meteorological encoder causes greater performance degradation in the NCP (117.1%) and NJP (157.3%) plain regions compared to the YTMR mountainous area (53.0%). In contrast, removal of the geographical encoder causes the most severe degradation in the YTMR (581.2%) among the three regions, particularly compared to the NJP, which shows only an 11.8% increase in RMSE.

Mechanism-level ablation experiments reveal that removing the site-type-aware fusion layer causes the most severe performance degradation across all regions, with RMSE increasing by 5103.7%, 2436.1%, and 1390.0% in the YTMR, NCP, and NJP, respectively. This demonstrates that the site-type-specific processing for natural, agricultural, and urban sites is crucial for model generalization across heterogeneous hydrogeological regions. Similarly, disabling the spatial attention module leads to substantial degradation, with RMSE increases exceeding 1000% in all three study regions (reaching 2591.3% in the YTMR, 1095.8% in the NCP, and 1420.9% in the NJP), which validates the essential role of adaptive spatial weighting based on four-dimensional similarity in capturing complex spatial dependencies in groundwater systems. Ablation of the dual-branch architecture within the historical encoder reveals distinct regional patterns; specifically, removing the CNN branch causes significantly greater performance degradation in the YTMR (1639.4%) than removing the LSTM branch (48.9%), whereas the opposite trend emerges in the NCP and NJP, where removing the LSTM branch causes greater degradation (618.4% and 540.3%, respectively) than removing the CNN branch (97.6% and 97.7%, respectively).”

Table 1. Summary of datasets across the three study regions. It would be great if the authors could report quantiles/percentiles instead of mean and standard deviations, that would give a much better overview of the sites.

Response

We sincerely appreciate the reviewer for this valuable suggestion. We have revised Table 1 to report the 25th percentile (Q_{25}), median (Q_{50}), and 75th percentile (Q_{75}) for all continuous variables, shown in parentheses below the mean ± standard deviation values.

Table 1. Summary of datasets across the three study regions. Values for continuous variables are presented as mean ± standard deviation, with the 25th percentile, median, and 75th percentile shown in parentheses below as (Q_{25}, Q_{50}, Q_{75}).

Attributes	NJP	NCP	YTMR
Record number	116864	58432	58432
Site numbers	64	32	32
Agricultural sites	23	19	14

Natural sites	17	5	11
Urban sites	24	8	7
Groundwater depth (m)	15.4±10.4 (4.8, 15.8, 24.3)	22.7±8.1 (16.3, 23.8, 28.4)	21.9±22.9 (4.8, 11.7, 33.0)
Rainfall (mm/day)	2.7±9.0 (0, 0, 0.9)	1.5±5.8 (0, 0, 0.2)	1.2±4.4 (0, 0, 0.3)
Average temperature (℃)	14.7±9.8 (5.9, 16.3, 23.5)	12.9±11.4 (2.1, 14.7, 23.6)	7.2±13.0 (-3.9, 9.0, 19.1)
Maximum temperature (℃)	19.5±9.8 (10.8, 21.4, 27.9)	18.5±11.5 (8.1, 20.9, 28.7)	14.0±13.3 (2.2, 16.2, 25.8)
Minimum temperature (℃)	10.8±10.0 (2.1, 11.5, 19.9)	8.0±11.4 (-2.6, 9.3, 18.8)	1.6±12.8 (-9.2, 2.6, 13.0)
Elevation (m)	13.2±11.9 (3.0, 10.0, 20.2)	32.0±14.6 (25.0, 32.0, 40.0)	776.8±358.7 (506.2, 673.5, 929.0)
Slope (º)	0.6±0.3 (0.4, 0.6, 0.8)	6.3±4.2 (3.2, 4.9, 8.2)	17.4±21.0 (6.0, 9.5, 21.7)
Watershed numbers	3	4	5
Aquifer types	2	2	2
Distance to the nearest river (m)	571.9±442.7 (247.0, 387.5, 889.0)	2162.1±2320.5 (681.0, 1400.0, 2600.0)	4406.3±8612.2 (392.0, 1126.0, 3025.0)
Population density (capita/km²)	503.2±164.6 (350.2, 491.0, 616.1)	1098.3±708.3 (657.2, 791.1, 901.4)	196.4±185.5 (80.7, 92.9, 377.5)
GDP (10^8 yuan)	2031.4±2252.2 (500.8, 707.5, 3601.3)	915.3±1367.5 (122.0, 200.4, 400.1)	147.1±28.1 (131.5, 150.3, 163.4)
Primary industry GDP (10^8 yuan)	206.1±200.4 (75.2, 97.4, 341.4)	104.3±154.5 (22.0, 24.6, 33.7)	32.9±23.3 (3.4, 36.2, 51.8)
Secondary industry GDP (10^8 yuan)	856.0±958.6 (203.2, 282.2, 1508.1)	309.9±483.0 (32.2, 65.4, 163.6)	42.9±20.7 (27.9, 33.6, 49.5)
Tertiary industry GDP (10^8 yuan)	969.8±1102.2 (236.0, 337.2, 1734.4)	492.1±735.8 (67.0, 85.6, 255.8)	71.3±28.3 (54.6, 63.6, 72.4)
Domestic water (L/(capita·d))	131.7±19.8 (121.7, 132.0, 140.2)	58.0±19.4 (42.2, 51.1, 71.2)	139.7±133.8 (83.2, 98.9, 153.6)
Industrial water ($m^3/(10^4$ yuan))	15.9±8.2 (11.0, 13.7, 16.1)	20.9±14.6 (9.9, 14.5, 31.1)	13.1±10.6 (4.7, 9.8, 19.1)
Irrigation water (m^3/mu)	405.1±69.6 (345.8, 406.1, 472.9)	120.8±56.9 (89.9, 125.9, 157.5)	21.7±10.7 (10.9, 21.9, 28.1)
Cropland in 500m buffer (%)	45.7±31.5 (15.6, 43.9, 73.4)	66.0±30.2 (42.8, 68.5, 94.9)	49.3±30.1 (28.4, 55.1, 69.3)
Forest in 500m buffer (%)	0 (0, 0, 0)	0.1±0.2 (0, 0, 0)	0.8±1.8 (0, 0, 0.5)
Grassland in 500m buffer (%)	0.1±0.2 (0, 0, 0)	0.1±0.3 (0, 0, 0)	9.0±14.8 (0, 0.5, 10.3)
Waters in 500m buffer (%)	3.2±8.99 (0, 0, 0.4)	0.1±0.2 (0, 0, 0)	1.3±4.7 (0, 0, 0)

Artificial surface in 500m buffer (%)	50.9±31.8 (23.4, 52.1, 73.9)	33.7±30.3 (3.5, 31.4, 57.1)	39.5±24.8 (22.4, 34.9, 54.8)

Fig 6: I think this is a great comparison with interesting insights, but especially in the top left panel it would be helpful to maybe present it as a box plot across groups, rather than a line plot.

Response

We sincerely appreciate the reviewer for this constructive suggestion. We have revised panel (a) of Figure 6 from a line plot to a grouped box plot, which displays the full distribution of MAPE for each site type (agricultural, natural, urban) under drought, normal conditions, and extreme rainfall in each of the three study regions. We also added discussion of the interquartile range and box width variations revealed by the box plots to provide a more comprehensive interpretation of model performance stability under different hydrological conditions. We have revised the sentences as follows:

"Compared to normal conditions, model prediction accuracy was substantially improved during drought events, with MAPE decreasing by 67.6% and 50.5% for agricultural and natural sites in the NCP, and by 51.3% for agricultural sites in the YTMR; these improvements were accompanied by narrower interquartile ranges, indicating more concentrated and consistent prediction errors across sites during such conditions. In contrast, prediction accuracy decreased during extreme rainfall events, with MAPE increasing by 41.9% for natural sites in the YTMR and by 38.6% for agricultural sites in the NJP compared to normal conditions. Notably, the model maintained relatively stable performance in the NCP under extreme rainfall, with MAPE increases of only 7.7% and 10.6% for agricultural and urban sites, respectively."

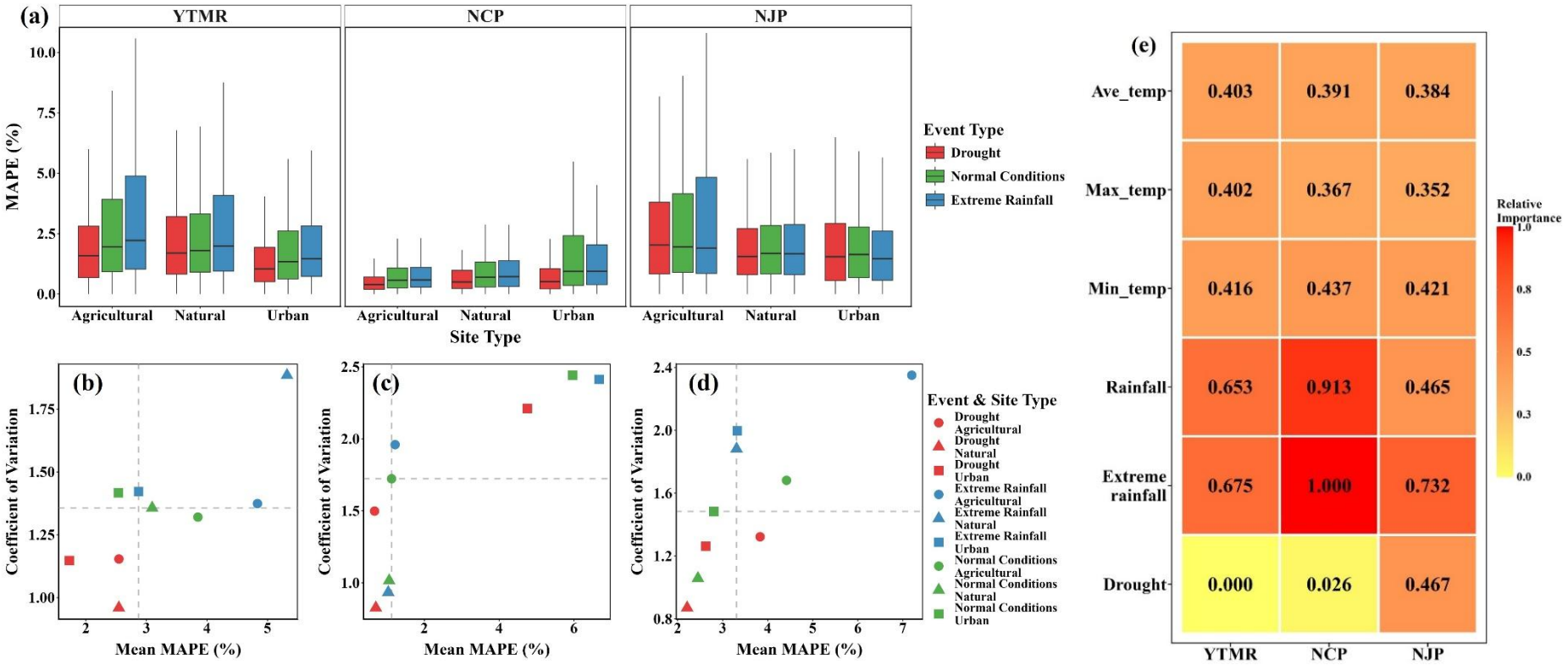

Figure 6. Model performance under droughts, normal conditions, and extreme rainfall. (a) Grouped box plots showing MAPE variation across droughts, normal conditions, and extreme rainfall for different site types in study regions; (b-d) scatter plots of coefficient of variation versus mean MAPE under droughts, normal conditions, and extreme rainfall in YTMR, NCP, and NJP, respectively; (e) heatmap displaying relative importance of meteorological variables and extreme events across the three study regions.

Reference

Asif, N.A., Sarker, Y., Chakrabortty, R.K., Ryan, M.J., Ahamed, M.H., Saha, D.K., Badal, F.R., Das, S.K., Ali, M.F. and Moyeen, S.I. (2021) Graph neural network: A comprehensive review on non-euclidean space. IEEE Access 9, 60588-60606.

Bai, T. and Tahmasebi, P. (2023) Graph neural network for groundwater level forecasting. Journal of Hydrology 616, 128792.

Han, K., Xiao, A., Wu, E., Guo, J., Xu, C. and Wang, Y. (2021) Transformer in transformer. Advances in neural information processing systems 34, 15908-15919.

Knoben, W.J., Freer, J.E. and Woods, R.A. (2019) Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. Hydrology and Earth System Sciences 23(10), 4323-4331.

Martinez-Ruiz, A., Gomez-Gil, P. and Fonseca-Delgado, R. (2025) An Analysis of Spatio-Temporal Graph Neural Networks Based on Synthetic Time Series with Known Structural Dependencies, pp. 313-324, Springer.

McCuen, R.H., Knight, Z. and Cutter, A.G. (2006) Evaluation of the Nash–Sutcliffe efficiency index. Journal of Hydrologic Engineering 11(6), 597-602.

Ohmer, M., Liesch, T., Habbel, B., Heudorfer, B., Gomez, M., Clos, P., Nölscher, M. and Broda, S. (2026) GEMS-GER: a machine learning benchmark dataset of long-term groundwater levels in Germany with meteorological forcings and site-specific environmental features. Earth Syst. Sci. Data 18(1), 77-95.

Riedmiller, M. and Lernen, A. (2014) Multi layer perceptron. Machine learning lab special lecture, University of Freiburg 24, 11-60.

Song, S., Xu, Y., Zhang, J., Li, G. and Wang, Y. (2016) The long-term water level dynamics during urbanization in plain catchment in Yangtze River Delta. Agricultural Water Management 174, 93-102.

Taccari, M.L., Wang, H., Nuttall, J., Chen, X. and Jimack, P.K. (2024) Spatial-temporal graph neural networks for groundwater data. Scientific Reports 14(1), 24564.

Zhang, K., Xie, X., Zhu, B., Meng, S. and Yao, Y. (2019) Unexpected groundwater recovery with decreasing agricultural irrigation in the Yellow River Basin. Agricultural Water Management 213, 858-867.

Zhao, Y., Yang, L., Pan, H., Li, Y., Shao, Y., Li, J. and Xie, X. (2025) Spatio-temporal prediction of groundwater vulnerability based on CNN-LSTM model with self-attention mechanism: A case study in Hetao Plain, northern China. Journal of Environmental Sciences 153, 128-142.