This manuscript develops a Hierarchical Hydrological Knowledge-guided Attention Network (HHA-Net) for 14-day groundwater-depth forecasting across three hydroclimatically distinct regions in China. The model integrates historical groundwater depth, meteorological conditions, geographical characteristics, and human activities through specialized encoders, site-type-aware feature fusion, and spatiotemporal attention mechanisms. A key methodological innovation is the four-dimensional spatial-similarity representation, which accounts for geographic distance, aquifer type, land-use-based site type, and watershed affiliation rather than relying solely on geographic proximity. The model is evaluated using 128 monitoring sites and more than 230,000 daily observations and is compared with several baseline models, including MLP, CNN-LSTM, Transformer, GNN, and ST-GNN. HHA-Net generally outperforms the selected baseline implementations. The main scientific finding is that the dominant predictive controls vary among regions, with geographical factors being most important in the mountainous region, human activities in the intensively exploited plain, and historical groundwater conditions in the humid coastal plain. Overall, the framework is methodologically innovative, and the large multi-regional dataset provides a valuable basis for evaluating groundwater prediction under contrasting hydrogeological and anthropogenic conditions.

Response

We sincerely appreciate the reviewer for the thorough and constructive review of our manuscript, and for the positive assessment of its methodological contribution and value. We have carefully addressed the reviewer's detailed comments below and revised the manuscript accordingly, and we believe these revisions have further strengthened the clarity and rigor of the manuscript.

My main comment is:

Line 266, 'All baseline models adopt identical training strategies and hyperparameter settings as HHANet to ensure fair comparative experiments.' This raises concern that the baseline models may not have been adequately optimized. The authors should clarify how each baseline was trained and tuned to demonstrate that HHA-Net's improved performance results from its framework rather than from comparison with poorly optimized baseline models.

Response

We sincerely appreciate the reviewer for this valuable and constructive comment. To address this concern about the baseline models may not have been adequately optimized, we have replaced the original uniform-hyperparameter setting with an independent hyperparameter optimization 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. 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. The code of the baseline models (run_models.py) has been made publicly available on Zenodo (https://doi.org/10.5281/zenodo.21915502).

Importantly, after model-specific optimization, HHA-Net continues to outperform the baseline models across the great majority of site types and regions, indicating that its predictive advantage from its architectural design. For example, the interquartile MAPE ranges of ST-GNN (3.13%–19.86%) and Transformer (1.94%–11.97%) each span more than sixfold across YTMR agricultural sites, whereas HHA-Net's range remains comparatively narrow (1.38%–4.49%). A similar pattern is observed at natural sites in the YTMR, where GNN (10.30%) and Transformer (7.18%) exhibit mean MAPE values three to four times higher than HHA-Net (2.50%), with wider, right-skewed quantile ranges (e.g., 2.54%–17.51% for GNN) compared to HHA-Net (1.29%–3.59%). At urban sites in the YTMR, the gap is even more pronounced, with MAPE (13.97%) of ST-GNN exceeding four times that of HHA-Net (3.34%). These results indicate that despite being individually optimized, baseline architectures without hydrological knowledge guidance still struggle to robustly capture groundwater dynamics under complex spatial and anthropogenic influences.

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.

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 (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**

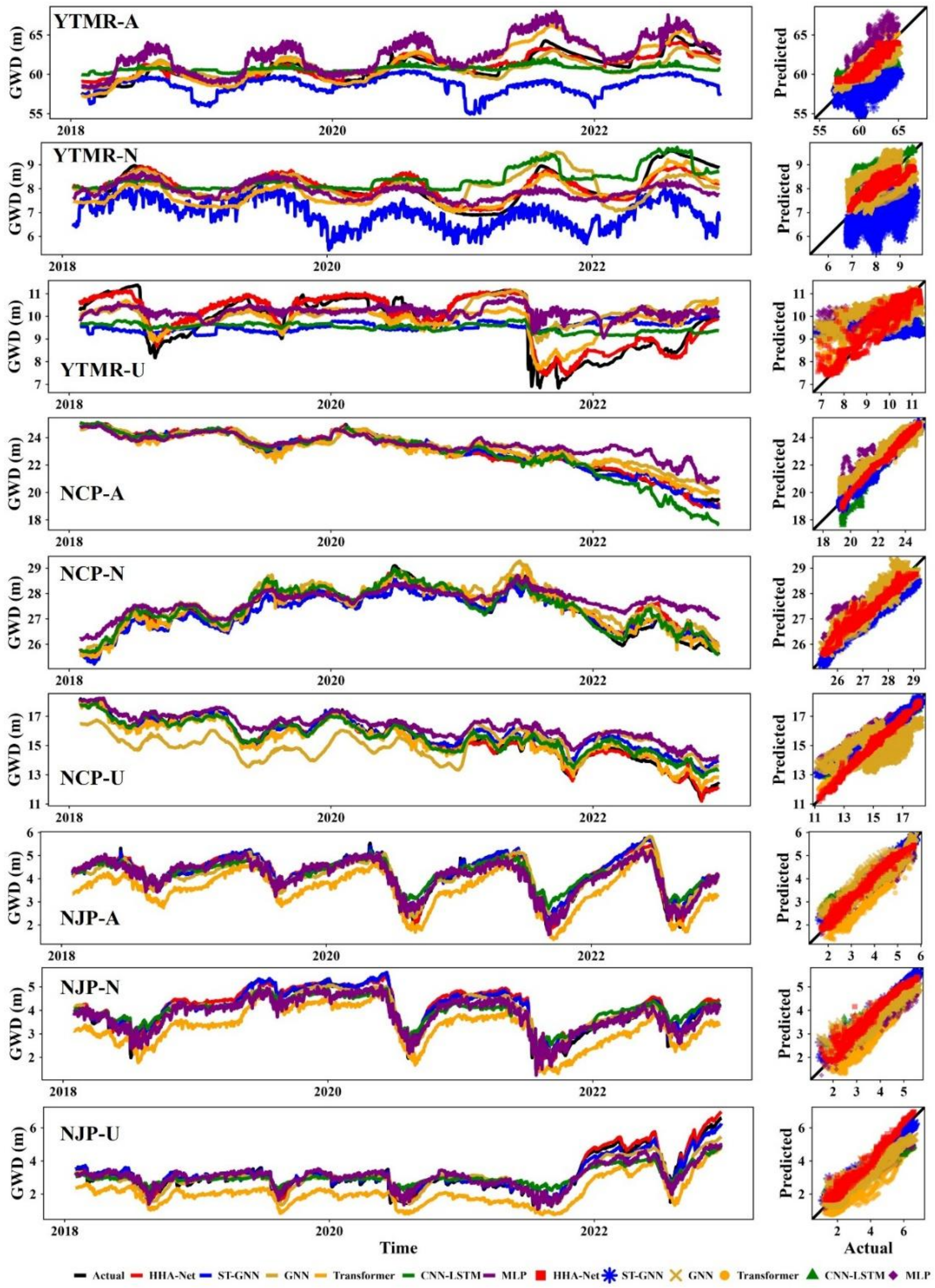

Figure 3: Comparison of GWD prediction performance across the three site types (agricultural (A), natural (N), and urban sites (U)) in the three study regions. Left panels show time series predictions for typical sites in YTMR, NCP, and NJP (prediction performance for all sites is shown in Figure S1,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

Table 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 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

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}
LYRP	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
	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**

Figures:
Figure 5 presents four performance metrics, but the discussion mainly highlights a few poorly performing sites without explaining differences among the spatial patterns. Some sites show low MAPE or RMSE but weak NSE or R2, suggesting that the model captures mean groundwater depth better than temporal variability. The authors should discuss these discrepancies and relate site-level performance to hydrogeological and land-use characteristics.

Response
We sincerely appreciate the reviewer for this constructive comment. In the revised manuscript, we have expanded the discussion of spatial performance patterns to explain low absolute-error metrics (MAPE and RMSE) and weak variance-normalized metrics (NSE and R^2), and to relate site-level performance to hydrogeological and land-use characteristics. Specifically, for the YTMR region, we clarify that natural sites in high-elevation mountainous zones (>850 m) show low MAPE and RMSE but weak NSE and R^2 because the thick vadose zone buffers meteorological variability, producing a near-stationary hydrograph with very low observed variance; since the NSE denominator scales with observed variance, even minor residuals are disproportionately amplified. For the NJP region, we add discussion about an urban site with highly impervious surface coverage (about 99.5% within a 500 m buffer), which similarly shows low MAPE and RMSE but weak NSE and R^2, attributed to surface sealing suppressing natural recharge and compressing the dynamic range of groundwater depth.
We have added the discussion and revised the sentences in the manuscript as follows:
"In terms of spatial distribution, the YTMR region shows relatively uniform model performance, except for two urban sites located in foothill regions and transition zones that exhibit slight degradation (RMSE up to 2.09 m), where complex fissure groundwater interactions and rapid lateral recharge pose challenges to model prediction. However, for natural sites in the high-elevation mountainous zones (elevations >850 m), the model achieves exceptionally low absolute errors (MAPE < 1.5% and RMSE < 0.35), yet yields low NSE and R^2 values (NSE < 0 and R^2 < 0.2). The thick vadose zone heavily buffers meteorological variability, resulting in an exceptionally low temporal dynamic range in the groundwater depth. 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). In the NJP region, model performance is consistently excellent, with spatial heterogeneity strongly mediated by inland-to-coastal hydrodynamics, except for one agricultural site in the western inland area that shows relatively higher errors (MAPE of 23.06%), which may be caused by episodic agricultural pumping from shallow aquifers. Notably, one urban site in

the NJP displays an error–goodness-of-fit decoupling similar to that in YTMR; situated in a highly impervious urban core (about 99.5% artificial surfaces within a 500 m buffer), it shows low absolute errors (MAPE = 0.66%, RMSE = 0.14 m) alongside weak NSE and R^2 (NSE = -0.09, R^2 = 0.15), which may be caused by extensive impermeable surfaces suppressing natural recharge and compressing the dynamic range of groundwater depth, thereby amplifying variance-normalized error metrics.”

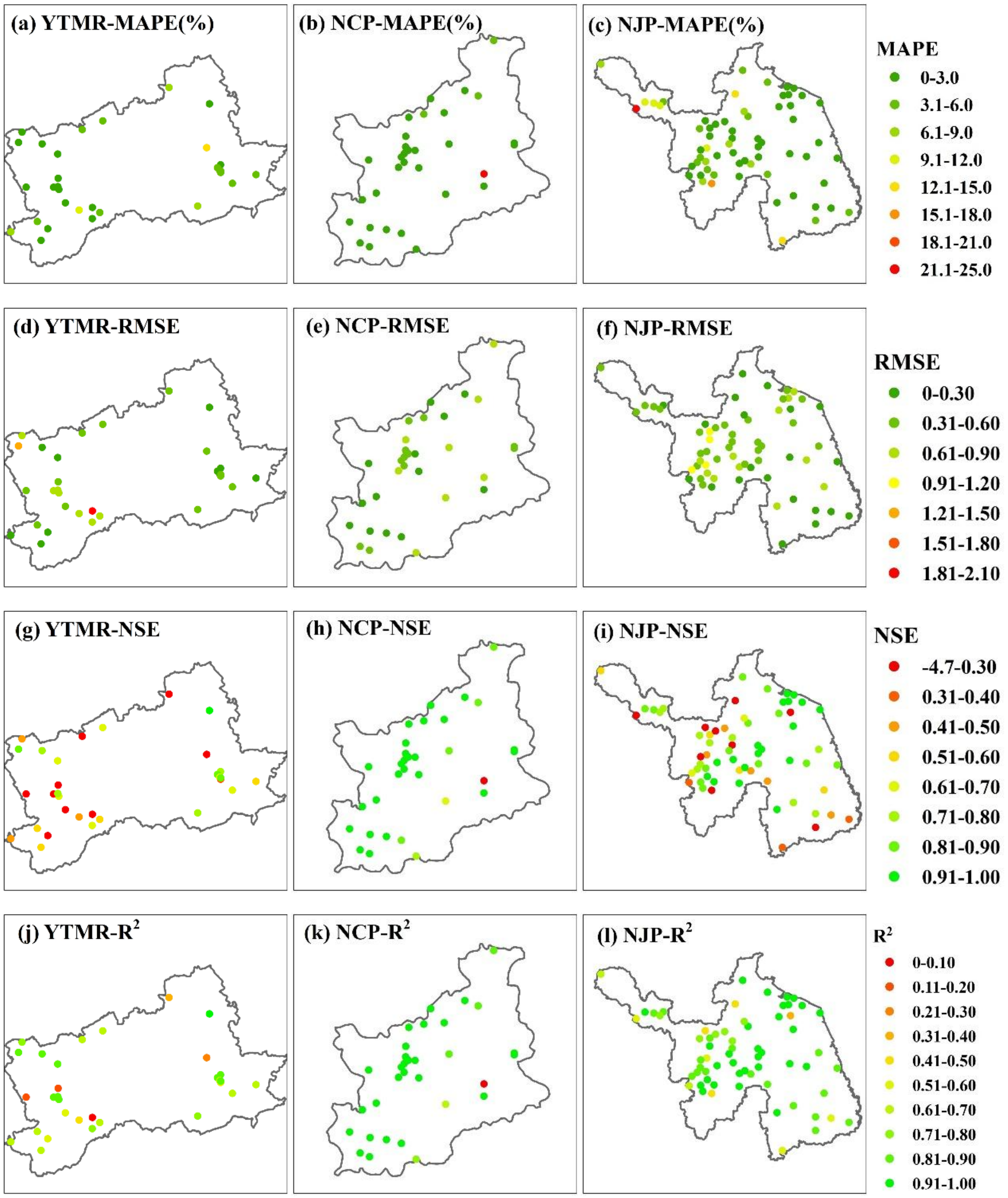

Figure 5: Spatial distribution of model performance metrics (MAPE (%), RMSE (m), NSE, and R^2) across groundwater monitoring sites in the three study regions.

Figure 6 lacks panel labels (a–d) and clear legends for panels (b–d), making it difficult to interpret. A simplified legend, direct point labels, or clearer panel separation would improve readability. Please also clarify whether all region–site-type combinations are shown and identify any excluded combinations.

Response

We sincerely appreciate the reviewer for these helpful suggestions. We have revised Figure 6 by adding clear panel labels and legends to all subplots, with each panel now enclosed by a distinct border to improve visual separation. For panels (b)–(d), we replaced the previously ambiguous and separated symbol-color legends with a unified, simplified legend that directly labels each point by its combined events (drought, extreme rainfall and normal conditions) and site types (agricultural, natural and urban sites), making the scatter plots much easier to interpret. We also confirm that all nine combinations of three site types and three event types are included with no exclusions, as each of panels (b)–(d) corresponds to one study region (YTMR, NCP, and NJP, respectively).

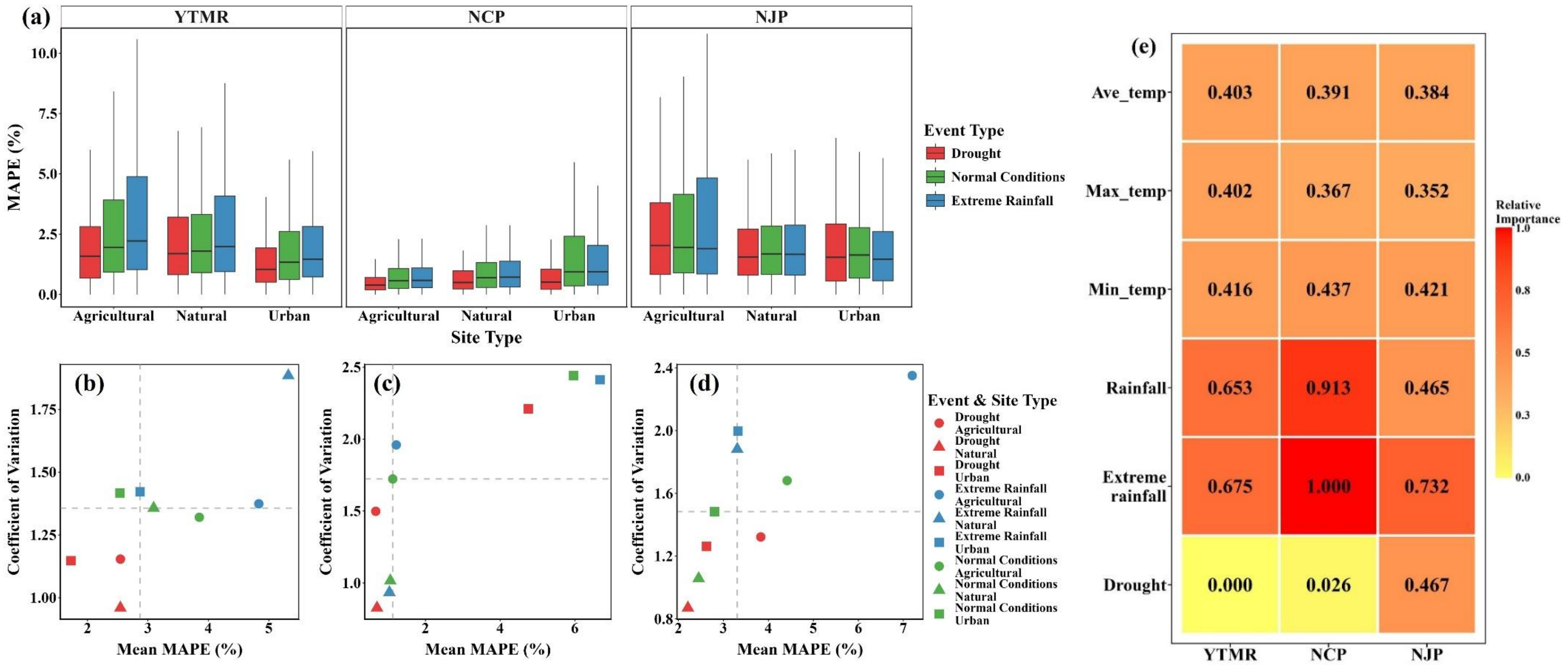

Figure 6: Model performance under droughts, normal conditions, and extreme rainfall. (a) Grouped box plots showing MAPE variation across droughts, normal conditions, and extreme rainfall for different site types in study regions; (b-d) scatter plots of coefficient of variation versus mean MAPE under 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.

Panels 7a and 7b present largely the same region- and site-type-specific bootstrap uncertainty information, with panel 7b showing a normalized radar-chart version of panel 7a. The authors should consider removing one panel or replacing it with an analysis that provides additional insight.

Response

We sincerely appreciate the reviewer for this valuable and constructive comment. To address this concern, we have replaced Panel 7b with a new relative uncertainty analysis, defined as the ratio of the Bootstrap Standard Error to the RMSE mean (SE/RMSE, expressed as %) for each region and site-type combination. Unlike Panel 7a, which reflects the absolute magnitude of uncertainty, this new panel evaluates uncertainty relative to the corresponding prediction error, allowing us to assess whether high absolute uncertainty in a given region is proportionally significant or simply a consequence of larger baseline prediction errors. This analysis confirms that high uncertainty in the NCP is not merely a scale effect but reflects a higher proportional uncertainty, while low uncertainty in the NJP is both absolutely and proportionally the most stable.

We have revised the sentences in the manuscript as follows: **"This study employs stratified bootstrap methods to analyze model uncertainty (Figure 7). The NCP shows the highest absolute and relative uncertainty, particularly at urban sites (BSE = 0.013; SE/RMSE = 1.9%), with uncertainty metrics increasing rapidly from T+1 to T+6 before remaining stable. The YTMR region, as a relatively natural mountainous area, demonstrates moderate uncertainty levels (BSE 0.003-0.011), and its relative uncertainty remains similarly moderate and balanced across site types (0.7-1.2%). The NJP region shows the lowest prediction uncertainty (0.004) and maintains consistently low and stable uncertainty throughout the multi-step predictions; correspondingly, its relative uncertainty is also the lowest and most stable among the three regions (0.65-0.72%)."**

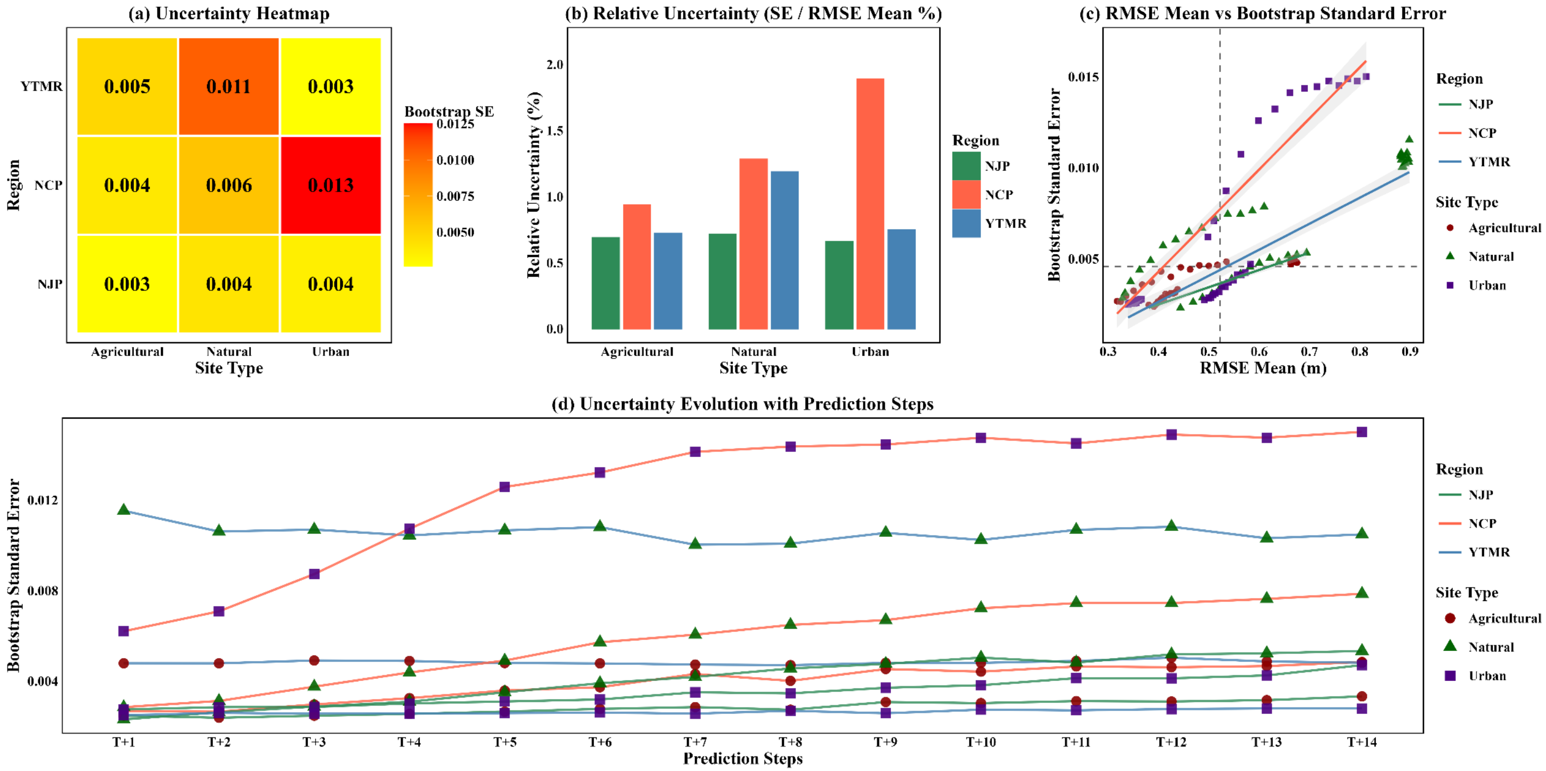

Figure 7: Uncertainty analysis across the three study regions and site types. (a) Uncertainty heatmap showing average bootstrap standard errors for different combinations of regions and site types. (b) Relative uncertainty bar chart displaying the percentage ratio of Bootstrap Standard Error (SE) to RMSE mean across site types for each region. (c) Scatter plot of RMSE mean estimates versus bootstrap standard errors, with region-specific regression lines and 95% confidence intervals (shaded areas). Dashed lines indicate median values for both axes, creating four quadrants for uncertainty categorization. (d) Temporal patterns of bootstrap standard errors across prediction steps from T+1 to T+14.

Figure 9 reports rainfall thresholds of 0.70, −0.16, and 0.07, but the manuscript does not clearly explain the rainfall normalization, the corresponding values in physical units, whether these thresholds are site-specific or regional averages, or how a negative normalized threshold should be interpreted. Referring to −0.16 as a "negative rainfall threshold" may therefore be misleading, as it could be mistaken for physically negative rainfall.

Response

We sincerely appreciate the reviewer for identifying this point of potential confusion. We have revised the manuscript to address this concern. In Appendix A, we added that rainfall is standardized via z-score normalization using each region's training-period mean and standard deviation before entering the threshold module, with the same statistics applied to the validation and test periods; the learned threshold τ is therefore dimensionless and expressed in this standardized space rather than as an absolute physical quantity. The corresponding physical value can be recovered by adding the product of τ and the training-period standard deviation to the training-period mean. The threshold is learned as a single shared parameter for each region. Regarding the negative threshold for the NCP, we clarify that this does not imply physically negative rainfall; rather, it means that even very small rainfall changes are sufficient to induce significant GWD fluctuations, without requiring exceedance of a minimum rainfall amount. We have added explanations in the manuscript as follows:

"The NCP exhibits a pronounced negative rainfall threshold (-0.16), indicating that the activation level already reaches 0.83 when rainfall is at the region's average condition (calculated by Equations A.3 and A.4). As a result, even very small rainfall changes are sufficient to induce significant GWD fluctuations, without requiring the exceedance of a minimum rainfall amount. This anomalous sensitivity reflects fundamental transformations of the aquifer system. Decades of over-exploitation have created extensive cones of groundwater depression in the area, leading to dramatically steeper hydraulic gradients compared with regions having adequate groundwater resources (Wang et al. 2021, Yuan et al. 2013)."

"To simulate nonlinear threshold response characteristics, we introduce learnable threshold parameters, whereby system responses are significantly amplified when the absolute values of input signals exceed the thresholds. The modulation factor M is defined as:

$$M = \mathrm{Sigmoid}[10 \times (x - \tau)] \quad \textbf{(A.3)}$$

$$\mathrm{Sigmoid}(z) = \frac{1}{1 + e^{-z}} \quad \textbf{(A.4)}$$

where x represents the meteorological input variables (temperature and rainfall in the study), τ is the response threshold parameter learned through model optimization, and the coefficient 10 serves as a scaling factor that steepens the sigmoid transition, enabling the model to better capture threshold-driven nonlinear responses. This formulation ensures that meteorological forcing has

significant effects on GWD only when input magnitudes surpass the response threshold.
The meteorological inputs are standardized through z-score normalization using the mean and standard deviation of training-period meteorological data before being passed into the threshold module in each region, and the same statistics are subsequently applied to the validation and test periods. Consequently, the response threshold learned by the model is expressed in this standardized space and is dimensionless, rather than an absolute physical quantity. For precipitation, the corresponding physical value can be recovered by adding the product of τ and the training-period standard deviation to the training-period mean. For each region, the threshold reported in this study is learned as a single shared parameter, reflecting the overall hydrogeological characteristics of that region. It should be emphasized that a negative threshold does not imply physically negative rainfall. Rather, it indicates that the response gate remains substantially open even under minimal rainfall, so that GWD fluctuations can be induced without requiring a minimum rainfall threshold to be exceeded."

Minor comments:
Line 435-440, the statement that the geographical encoder caused the largest performance degradation in the YTMR is inconsistent with the reported ablation results. The historical encoder removal produced a 2213.2% increase in RMSE, compared with 581.2% for the geographical encoder. Please correct the inconsistency.

Response
We sincerely appreciate the reviewer for the careful reading and for pointing out this potential inconsistency, which helped us improve the clarity and precision of this discussion. We would like to clarify that the original statement was intended as a cross-regional comparison for the geographical encoder specifically, rather than a comparison across the four different encoders within the YTMR. We have revised the sentence to explicitly clarify that the comparison is made across regions for the same encoder, as follows: **"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."**

Line 449-450, the statement is incorrect for the NJP. Removing the CNN branch caused greater degradation than removing the LSTM branch only in the YTMR (1639.4% versus 48.9%). But in the NJP, removing the LSTM branch caused the larger degradation (540.3% versus 97.7%).

Response
We sincerely appreciate the reviewer for the careful review and for pointing out this incorrect statement. We have revised the sentences in the manuscript as follows: **"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).”

Lines 682 and 752, please be consistent on whether the maximum lag is 5 days or 15 days.

Response
We sincerely appreciate the reviewer for the careful review and for pointing out this inconsistency. We have corrected the corresponding sentence in Section 5.4 (Limitations and future work) to be consistent with Appendix A1, as follows: **“For example, the maximum lag time for groundwater response (15 days) is based on empirical synthesis from regional studies and may not be universally applicable across different hydrogeological environments or extreme conditions.”**

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.

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.

Wang, Y., Chen, Z., Chen, J. and Wei, W. (2021) Effects of ultrafiltration on salt migration and isotopic composition of pore water in groundwater depression area in Hengshui, NCP. Hydrological Processes 35(5), e14201.

Yuan, R., Song, X., Han, D., Zhang, L. and Wang, S. (2013) Upward recharge through groundwater depression cone in piedmont plain of North China Plain. Journal of Hydrology 500, 1-11.

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.