

Bayesian Optimization-based STA-BiLSTM Groundwater Level Prediction Model and Its Application in Beijing-Tianjin Area

Xiaoya Zhang¹, Xin Liao^{2,3,4}, Yun Shi^{2,3,4}, Yonghong Xue^{1,a,*}

¹*School of Science, University of Emergency Management, Beijing, 101601, China*

²*School of Environmental and Disaster Management, University of Emergency Management, Beijing, 101601, China*

³*Hebei Key Laboratory of Resource and Environmental Disaster Mechanism and Risk Monitoring, Sanhe, 065201, China*

⁴*Hebei Key Laboratory of Earthquake Dynamics, Sanhe, 065201, China*

^a*aristotle@ncist.edu.cn*

**Corresponding author*

Keywords: Bayesian optimization; BiLSTM; Groundwater level prediction; Spatio-temporal attention mechanism

Abstract: Groundwater level dynamics are shaped by the coupled influences of multiple hydrogeological factors and human activities, exhibiting pronounced nonlinearity and spatiotemporal heterogeneity. To address the limitations of traditional models in effectively capturing spatiotemporal features and parameter optimization difficulties, this study proposes a Bayesian optimization-based spatiotemporal attention bidirectional long short-term memory model (STA-BiLSTM). The proposed model integrates adaptive spatial neighborhood construction with a hybrid spatiotemporal attention mechanism, effectively fusing spatial correlations among multiple monitoring wells with temporal dependencies across time steps. Furthermore, Bayesian optimization is employed for global, automated hyperparameter and structural parameter tuning. To gain deeper insight into the hydrophysical significance underlying the model's internal mechanisms, experiments were conducted in the Beijing–Tianjin region, selecting groundwater-level rising and declining zones with markedly different evolutionary characteristics. Results demonstrate that the STA-BiLSTM model achieves higher predictive accuracy than comparative models under both evolution patterns. Compared with the baseline LSTM model, the root mean square error (RMSE) decreased by 38.4% and 7.9%, and the coefficient of determination (R^2) increased to 0.945 and 0.940 for the Beijing and Tianjin wells, respectively.

1. Introduction

Groundwater serves as a vital strategic resource supporting urban ecological sustainability and socioeconomic development [1]. In water-scarce regions such as the North China Plain, groundwater dynamics are jointly influenced by natural hydrogeological processes and intensive

human activities, leading to complex spatiotemporal evolution patterns [2]. In recent years, large-scale human interventions, including the South-to-North Water Diversion Project, have created a striking contrast in the Beijing–Tianjin region: groundwater levels are rebounding in the upstream Beijing Plain while continuing to decline in the downstream Tianjin Coastal Plain [3, 4, 5]. The distinct hydrogeological conditions determine different dominant drivers: groundwater recovery in Beijing is mainly governed by artificial recharge and river infiltration, while the decline in Tianjin is driven by deep groundwater overexploitation and limited recharge. This polarized pattern within the same hydrological system provides an ideal testbed for evaluating the adaptability and robustness of predictive models under contrasting hydrodynamic mechanisms.

In recent years, deep learning techniques—represented by recurrent neural networks (RNNs) and their advanced variant, long short-term memory (LSTM) networks—have been widely applied in hydrological modeling and have gradually become mainstream tools for groundwater level prediction [6-8]. Studies have shown that LSTM networks exhibit remarkable advantages in capturing long-term dependencies within complex hydrological time series [9,10]. Whether applied to multivariate modeling that integrates meteorological and anthropogenic factors or univariate forecasts of confined aquifer water levels, LSTM models consistently outperform traditional approaches [7, 11, 12]. To further enhance the model's ability to focus on critical information, researchers have introduced the attention mechanism into hydrological forecasting. By dynamically assigning weights to input features, attention mechanisms enable models to more effectively focus on key driving factors such as precipitation and groundwater extraction, thereby improving prediction stability and accuracy [13, 14]. Incorporating attention mechanisms into LSTM architectures also enhances model interpretability [15, 16]. Moreover, recent studies have attempted to develop unified spatiotemporal attention frameworks capable of simultaneously capturing spatial dependencies across well networks and temporal dynamic patterns, achieving regionally high-accuracy groundwater level predictions [17]. However, the above models based on unidirectional LSTM architectures can only utilize historical information. To fully exploit the bidirectional dependencies within time series, bidirectional long short-term memory (BiLSTM) networks have been introduced into hydrological prediction [18]. Their bidirectional propagation mechanism enables simultaneous learning of forward and backward sequence features, thereby enhancing the model's capability to represent complex nonlinear dynamic processes.

Despite the remarkable progress of existing deep learning models in groundwater level forecasting, two critical issues remain to be addressed. The first concerns model structure optimization. Most current studies, when dealing with multi-well monitoring data characterized by strong spatial correlations, tend to overlook the hydraulic connectivity among neighboring wells [19], resulting in insufficient utilization of spatial information. Although attention mechanisms have been proven effective, how to adaptively integrate spatial information from neighboring wells at varying distances with temporal dependencies from different historical time steps within multi-well monitoring networks remains an open challenge. The second issue involves the arbitrariness in hyperparameter determination. The performance of deep learning models is highly sensitive to the configuration of hyperparameters such as time window length, number of hidden units, and learning rate [20]. Existing studies typically rely on grid search [21] or random search [22] for hyperparameter tuning, both of which are inefficient and prone to suboptimal solutions in high-dimensional parameter spaces, thus limiting model performance. Bayesian optimization, as an efficient global optimization approach based on probabilistic surrogate models, can determine the optimal solution of complex objective functions with a limited number of iterations, thereby offering a powerful tool to address the hyperparameter optimization problem in deep learning models [23, 24].

Therefore, this study proposes a Bayesian optimization-based spatiotemporal attention

bidirectional long short-term memory prediction model. To address the challenge of limited cross-scenario generalization capability in existing models, a partitioned validation strategy is employed, conducting experiments under two representative hydrogeological conditions. The study aims to achieve three primary objectives: (1) Develop a BiLSTM architecture integrating dual attention mechanisms to effectively capture both the temporal dependency characteristics of groundwater level variations and the spatial correlation characteristics among multiple monitoring wells; (2) Implement Bayesian optimization for global automatic tuning of both hyperparameters and model structure, thereby enhancing prediction accuracy and automation efficiency; (3) Leverage the contrasting groundwater evolution patterns in Beijing (rising) and Tianjin (declining) to systematically evaluate the model's prediction precision and generalization capability across diverse hydrogeological evolution contexts, while exploring the underlying relationships between model internal mechanisms and hydrogeological processes.

2. Model Construction

2.1. BiLSTM

The long short-term memory (LSTM) network is a specialized variant of recurrent neural networks (RNN), first proposed by Hochreiter and Schmidhuber [25]. It was designed to address the vanishing and exploding gradient problems encountered by standard RNNs when processing long sequence data, thereby effectively capturing and learning long-term dependencies within sequences.

The bi-directional long short-term memory (BiLSTM) network combines two LSTM layers: a forward LSTM that processes the input sequence in chronological order, and a backward LSTM that processes it in reverse order [26](as shown in Figure 1). While LSTM excels at capturing long-term dependencies, BiLSTM further strengthens this capability by enabling the model to comprehend information both prior to and subsequent to a specific moment. This allows for a more comprehensive excavation of dynamic patterns in time series, offering superior fitting capabilities for groundwater level data characterized by distinct seasonality and lag effects [27].

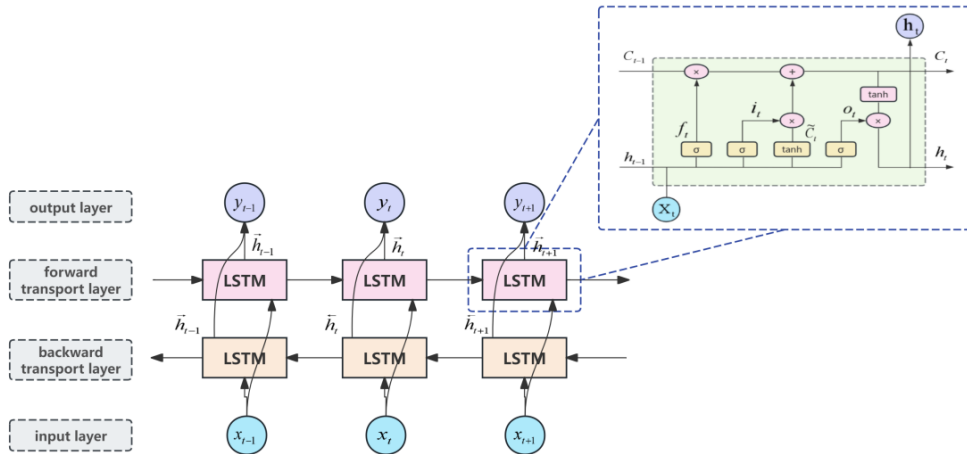


Figure 1: Architecture of the BiLSTM model.

2.2. Attention Mechanism

The attention mechanism is commonly used to optimize sequence processing in models. Its core idea is to enable the model to mimic the selective attention behavior of human cognition when

processing information; that is, when handling complex data, the model automatically assigns weights to focus on the parts most critical to the current task [28]. The attention mechanism is formulated as follows:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

where Q represents the query vector, which is the vector for which information is currently being sought; K represents the key vector, which is the set of reference vectors available for querying; V represents the value vector, containing the actual information corresponding to the keys, used for weighted summation to obtain the final output; d_k is the dimension of the key vector; QK^T calculates the correlation between the query and key vectors to obtain an attention score matrix; and Softmax is used to normalize the correlation scores into a probability distribution.

2.3. Bayesian Optimization Algorithm

Bayesian optimization is an efficient sequential decision-making strategy designed to solve global optimization problems for black-box functions [23, 29]. Its core philosophy utilizes a Gaussian Process (GP) [30] as a prior model to fit the objective function. GP models provide both predictive means for unobserved parameter points and quantified uncertainty estimates, furnishing the full posterior distribution essential for optimization guidance. Coupled with an acquisition function, this mechanism systematically identifies the next evaluation point to converge toward the global optimum [31].

To achieve automatic hyperparameter tuning and enhance predictive performance, this paper incorporates the Bayesian optimization algorithm into the STA-BiLSTM model. The algorithmic procedure includes the following steps: (1) Randomly select several initial sample points within the given parameter search space, calculate their corresponding objective function values, and construct an initial observation set; (2) Fit or update the Gaussian Process surrogate model based on the current observation dataset to approximate the distribution shape of the true objective function; (3) Maximize the acquisition function to determine the next hyperparameter combination with the highest expected improvement; (4) Evaluate the objective function using the true model and incorporate the new input-output pair into the observation set; (5) Repeat the iteration process until the preset number of iterations is reached or convergence conditions are met, ultimately outputting the global optimal parameter solution.

2.4. Spatio-Temporal Attention-BiLSTM Model

This paper constructs a spatiotemporal attention bidirectional long short-term memory (STA-BiLSTM) model incorporating Bayesian optimization. As illustrated in Figure 2, the STA-BiLSTM model comprises the following key components:

(1) **Spatial Neighborhood Construction:** Drawing upon spatial correlation analysis, an initial search radius R is defined, and all monitoring wells falling within range R of the target point are incorporated into the neighborhood set. To mitigate subjectivity in radius selection, a Bayesian optimization algorithm is introduced to automatically tune this key parameter. A spatial radius search algorithm is utilized to screen the set of neighboring monitoring wells for the target well, constructing a multivariate input matrix comprising historical water level data from both the target and neighboring wells.

(2) **Spatiotemporal Attention Feature Enhancement:** To address the varying degrees of influence that different neighboring monitoring wells have on the water level of the target well, the model calculates spatial weights for different sites via a spatial attention mechanism. This mechanism

adaptively identifies and assigns higher weights to strongly correlated sites while suppressing interference from weakly correlated ones, thereby yielding a spatially weighted feature sequence. Meanwhile, the temporal attention mechanism handles information in the time dimension; by computing importance weights for each time step in the input sequence, the model focuses on critical historical periods that significantly impact the current prediction result. Based on the results of Bayesian optimization, the sequence of the spatial and temporal attention modules can be adaptively adjusted (either spatial-first or temporal-first) to align with the optimal feature extraction path for different sites. This order selection is implemented through a conditional branching structure, enabling the model to prioritize either spatial or temporal information according to the dominant characteristics of the hydrogeological setting.

(3) Bidirectional Temporal Deep Modeling: The high-level feature sequences, having been doubly weighted and reconstructed by the spatiotemporal attention mechanisms, are fed into the BiLSTM layer. Leveraging its unique bidirectional recurrent structure, the BiLSTM simultaneously processes forward evolutionary information and backward dependency information of the sequence, fully exploiting the long-term memory and deep non-linear patterns within the dynamic changes of groundwater levels.

(4) Prediction Output: The hidden states output by the BiLSTM layer undergo linear mapping via a fully connected layer, ultimately generating the predicted groundwater level at the next time step.

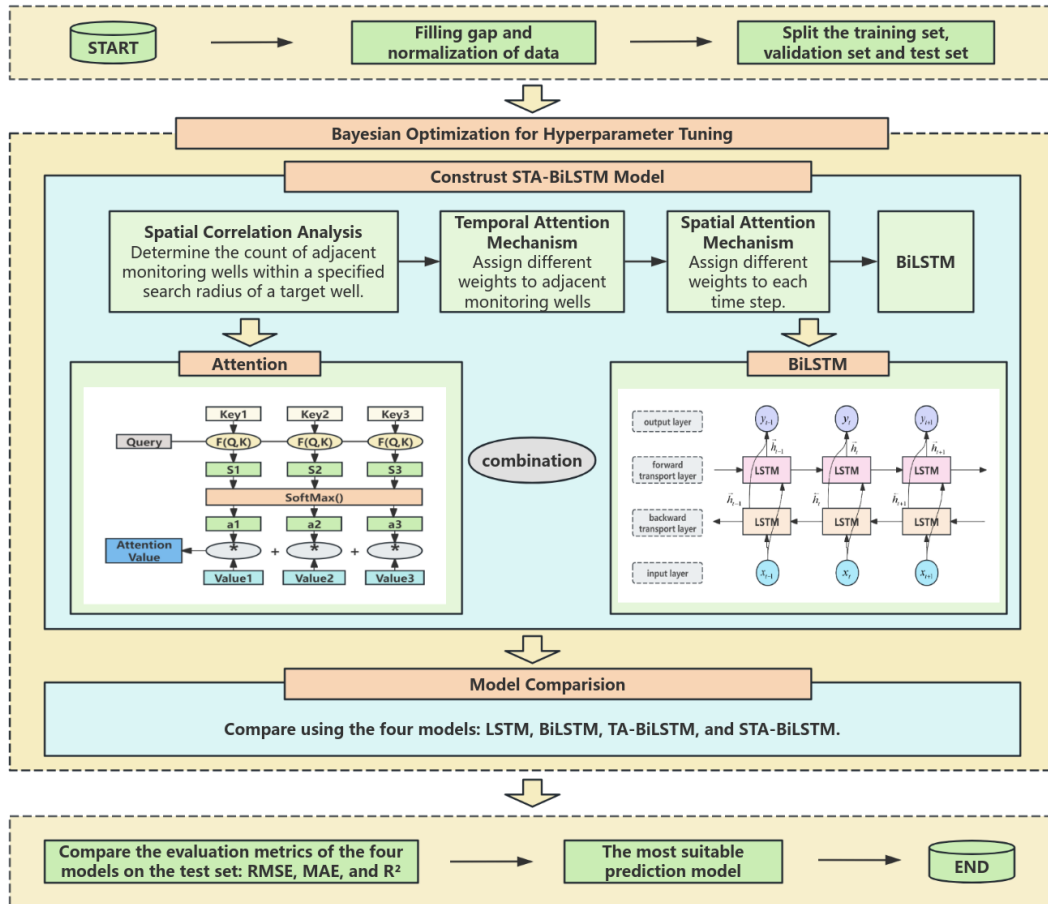


Figure 2: Workflow of the groundwater level prediction model.

3. Study Area and Data Sources

3.1. Overview of the Study Area

To ensure data quality for model training and to validate its generalization capability under complex hydrogeological conditions, 14 representative monitoring wells were selected based on data continuity and differing groundwater evolution patterns. Given the pronounced regional differentiation in groundwater level dynamics within the Beijing–Tianjin area, two validation scenarios were established. The first group comprises six monitoring wells in Beijing (BJ-1 to BJ-6), which exhibit continuous groundwater recovery trends, while the second group includes eight wells in Tianjin (TJ-1 to TJ-8), showing overall declining trends. These contrasting scenarios provide ideal experimental conditions for comprehensively evaluating the model's applicability under different groundwater evolution regimes.

3.2. Data Sources and Preprocessing

The groundwater depth monitoring data used in this study were obtained from the GeoCloud platform of the China Geological Survey (<https://geocloudsso.cgs.gov.cn/>). The data span from January 2018 to November 2024 with an hourly monitoring frequency, covering multiple monitoring points in the Beijing-Tianjin region. The training effectiveness and predictive accuracy of groundwater level models are highly dependent on the quality of the input data; therefore, input data must undergo preprocessing prior to prediction. To ensure the continuity and accuracy of the model input, this paper employs time-based interpolation to fill missing values, thereby obtaining complete time-series data. Furthermore, since there are significant differences in groundwater depths across various monitoring points and inconsistent magnitudes, directly inputting raw data into the model would lead to training instability. Consequently, the Min-Max normalization method [32] is adopted to standardize the data, mapping the original sequences to the $[0,1]$ interval:

$$y_{\text{norm}} = \frac{y_i - \min(y_i)}{\max(y_i) - \min(y_i)} \quad (2)$$

Where y_{norm} is the normalized value, y_i is the original value, $\min(y_i)$ is the minimum value of the dataset, and $\max(y_i)$ is the maximum value of the dataset. After the prediction results are output, inverse normalization is performed to restore the actual groundwater level values.

4. Groundwater Prediction and Evaluation in the Study Area

4.1. Dataset Partitioning

To evaluate the model's generalization capability, the time-series samples were partitioned into training, validation, and test sets at a ratio of 7:2:1. The training set was used for model parameter learning, the validation set for hyperparameter tuning and preventing overfitting, and the test set for final performance assessment. The specific time periods were as follows: training set from January 1, 2018, to November 4, 2022; validation set from November 4, 2022, to March 23, 2024; and test set from March 23, 2024, to November 30, 2024.

To fully account for the temporal dependency of groundwater levels, a sliding window method was employed for sample construction: given an input time step length T , the input features consist of the groundwater level sequence from the preceding T moments $[h_{t-T+1}, \dots, h_t]$, and the output corresponds to the level at the next moment h_{t+1} .

4.2. Model Settings and Hyperparameter Optimization

During the model training phase, the Adam optimizer was employed with mean squared error (MSE) as the loss function. An early stopping mechanism was utilized to prevent overfitting; training was automatically terminated if the validation set error did not significantly decrease for 20 consecutive epochs, retaining the model weights with the minimum validation error. The batch size was set to 32, and the maximum number of epochs was 200. Training and validation error curves were output during training to monitor convergence in real time, ensuring model stability and generalization. The validation phase used the validation set for interim performance assessment, providing a basis for subsequent hyperparameter fine-tuning.

Model performance is highly sensitive to hyperparameter configuration [33]. To further enhance model efficiency and predictive capability, this study employed Bayesian optimization for adaptive tuning of key hyperparameters. As shown in Table 1, the optimized parameters include time step size, number of hidden units, learning rate, dropout rate, and spatial radius. In addition, an attention order hyperparameter was introduced to adaptively determine the execution sequence of temporal and spatial attention mechanisms, enabling the model to learn the optimal feature extraction paths for different monitoring sites.

Table 1: Hyperparameter settings.

Parameter Name	Value Range	Purpose
Time Step (T)	(3,30)	Prediction Accuracy (Minimize MSE)
Hidden Units	(16,256)	Capture complex features
Learning Rate	(1e-5,1e-2)	convergence rate of model
Dropout Rate	(0.1,0.5)	Overfitting Prevention
Spatial Radius (R)	(0,10)	Neighborhood Determination Accuracy
Attention Order	(0, 1)	Selection Order

4.3. Model Evaluation Metrics

This study employs three distinct evaluation metrics-root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2)-to assess the predictive performance of the model. Each metric measures the discrepancy between predicted and actual values in different ways.

The calculation formulas for each metric are as follows:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}, \quad (3)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|, \quad (4)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad (5)$$

5. Results and Discussion

To systematically evaluate the predictive performance and generalization capability of the STA-BiLSTM model under different groundwater evolution regimes, this section presents a comprehensive analysis of the prediction results for Beijing (groundwater recovery zone) and Tianjin (groundwater decline zone). The analysis includes examination of the spatiotemporal parameters determined by Bayesian optimization and their hydrogeological implications,

comparison of the overall predictive accuracy of STA-BiLSTM against multiple benchmark models, visualization of typical monitoring wells to assess the model's ability to capture temporal details, and an in-depth discussion of model robustness together with the physical interpretation of the results.

5.1. Analysis of Spatiotemporal Parameter Characteristics

Table 2 summarizes the key spatiotemporal parameters automatically optimized for each monitoring well using the Bayesian optimization algorithm, including the optimal spatial search radius, the effective radius used, the number of neighboring wells, and the execution order of the attention modules. These parameters are not fixed but are adaptively generated according to the local hydrogeological conditions and monitoring network density at each site, thereby revealing the linkage between the model's internal mechanisms and the external environmental setting.

Table 2: Spatiotemporal parameters for each monitoring well.

Station	Optimal radius	Actual radius	Number of neighboring wells	Attention order ^a
BJ-1	3.6636	3.6636	1	0
BJ-2	0.0000	0.1000	4	1
BJ-3	1.7060	1.7060	3	1
BJ-4	0.1215	1.1215	1	1
BJ-5	3.6636	5.6636	1	0
BJ-6	0.0000	0.1000	3	1
TJ-1	0.8502	2.8502	2	1
TJ-2	6.9957	6.9957	6	0
TJ-3	0.4316	7.4316	3	1
TJ-4	0.000	5.1000	2	0
TJ-5	0.000	4.1000	1	0
TJ-6	9.1293	9.1293	9	1
TJ-7	10.0000	10.0000	10	0
TJ-8	10.0000	10.0000	8	0

^a Attention order: 0 indicates temporal before spatial attention; 1 indicates before temporal attention.

The Bayesian-optimized optimal radius represents the statistically optimal value determined by minimizing MSE on the validation set, reflecting the effective range of spatial information utilization by the model—termed the statistical effective neighborhood. Data distribution and noise levels may influence this radius, potentially causing it to differ from the hydrogeological influence radius, although the two may exhibit certain correlations. We conducted correlation analysis between this radius and the hydrogeological background to explore its underlying physical significance.

To accurately assess the model's strategy for utilizing spatial information, it is essential to distinguish between the optimal radius and the actual radius employed. The adaptive optimization process of this parameter essentially quantifies the model's learning of differences in mass transfer mechanisms across distinct hydrogeological units. When the optimal radius equals the actual radius, it indicates that the Bayesian-recommended search range provides sufficient spatial inputs, requiring no further expansion to leverage spatial correlations—as observed in wells BJ-3, TJ-2, and TJ-6. Conversely, when the optimal radius is substantially smaller than the actual radius, reliance solely on the recommended radius fails to meet the input requirements for neighboring wells, necessitating adaptive expansion to incorporate available adjacent wells, such as TJ-3, TJ-4, and TJ-5.

Table 2 reveals distinct regional patterns in the statistical effective neighborhood radii. Beijing wells generally exhibit smaller radii (mostly < 4 km), with some wells showing extremely small radii and dense neighboring wells, consistent with the high permeability of piedmont alluvial-proluvial fan aquifers, close hydraulic connectivity, and dense local well networks in the Beijing Plain. Tianjin wells, by contrast, display consistently larger radii, with TJ-7 and TJ-8 reaching the 10 km upper limit and incorporating numerous neighboring wells. This reflects the need for broader regional information to characterize slow regional flow fields under low-permeability conditions in the coastal plain, compounded by sparser monitoring well distribution. The selection of attention order also exhibits clear regional differentiation. Beijing wells predominantly adopt a spatial-first, temporal-second sequence, where rapid propagation of water level changes from neighboring wells makes spatial information the primary entry point for understanding dynamic processes. This preference aligns with the higher monitoring network density, enhancing spatial data availability. Tianjin wells show more varied patterns, but wells with pronounced declining trends (e.g., TJ-2, TJ-7) favor temporal-first, spatial-second. This likely stems from the dominance of regional overexploitation in deep confined aquifer systems, where long-term monotonic decline trends in time series better explain current levels than local spatial fluctuations, further exacerbated by sparse well networks and limited spatial data.

5.2. Comparison of Overall Prediction Performance among Different Models

Table 3 summarizes the predictive performance of the four models on the test sets for both Beijing and Tianjin monitoring wells. Figure 3(a-f) presents a bar chart comparison of the evaluation metrics for individual monitoring wells in the Beijing and Tianjin regions. Overall, the STA-BiLSTM model achieved optimal performance across all three evaluation metrics: RMSE, MAE, and R^2 .

Table 3: Performance evaluation metrics for each model.

Region	Performance	LSTM	BiLSTM	TA-BiLSTM	STA-BiLSTM
Beijing(BJ)	RMSE(m)	0.297	0.252	0.204	0.183
	MAE(m)	0.209	0.188	0.143	0.125
	R^2	0.890	0.912	0.941	0.945
Tianjin(TJ)	RMSE(m)	0.265	0.274	0.252	0.244
	MAE(m)	0.169	0.178	0.160	0.153
	R^2	0.928	0.912	0.932	0.940

For the Beijing wells, the STA-BiLSTM model demonstrates a significant performance advantage. Compared to the benchmark LSTM model, the STA-BiLSTM model reduces the average RMSE by 38.4% (from 0.297 m to 0.183 m) and the MAE by 40.2%, while improving the R^2 from 0.890 to 0.945. Analysis of the bar charts for Beijing wells reveals that BiLSTM shows a certain degree of improvement over LSTM, suggesting that the bidirectional structure enhances the capture of forward and backward dependencies within groundwater level rising trends. The further enhancement observed in the TA-BiLSTM model confirms the effectiveness of the temporal attention mechanism in identifying critical time steps. The superior performance of the STA-BiLSTM model is attributed to the incorporation of spatial information. Particularly at sites with multiple neighboring wells, such as BJ-2 and BJ-3, the spatial attention mechanism effectively corrects prediction biases inherent in single-sequence models, allowing the model to maintain extremely high fitting accuracy even in non-stationary rising sequences.

Although the overall performance differences among models were slightly smaller for the Tianjin wells compared to Beijing, the STA-BiLSTM model still maintained optimal performance.

Its average RMSE was 0.244 m, a 7.9% reduction compared to LSTM (0.265 m), with an R^2 reaching 0.940. Notably, the performance of the BiLSTM model in the Tianjin region was slightly inferior to that of the unidirectional LSTM (RMSE 0.274m vs 0.265m). This may stem from the strong monotonicity of the sequences caused by the long-term decline in Tianjin's groundwater levels, where the backward propagation of the bidirectional structure failed to provide additional gain and may have instead introduced noise. However, the STA-BiLSTM model overcame this limitation through the introduction of the spatiotemporal attention mechanism. At sites with numerous neighboring wells, such as TJ-6 and TJ-7, the adaptive allocation of spatial weights facilitated the effective aggregation of regional water level decline information, thereby achieving improved accuracy.

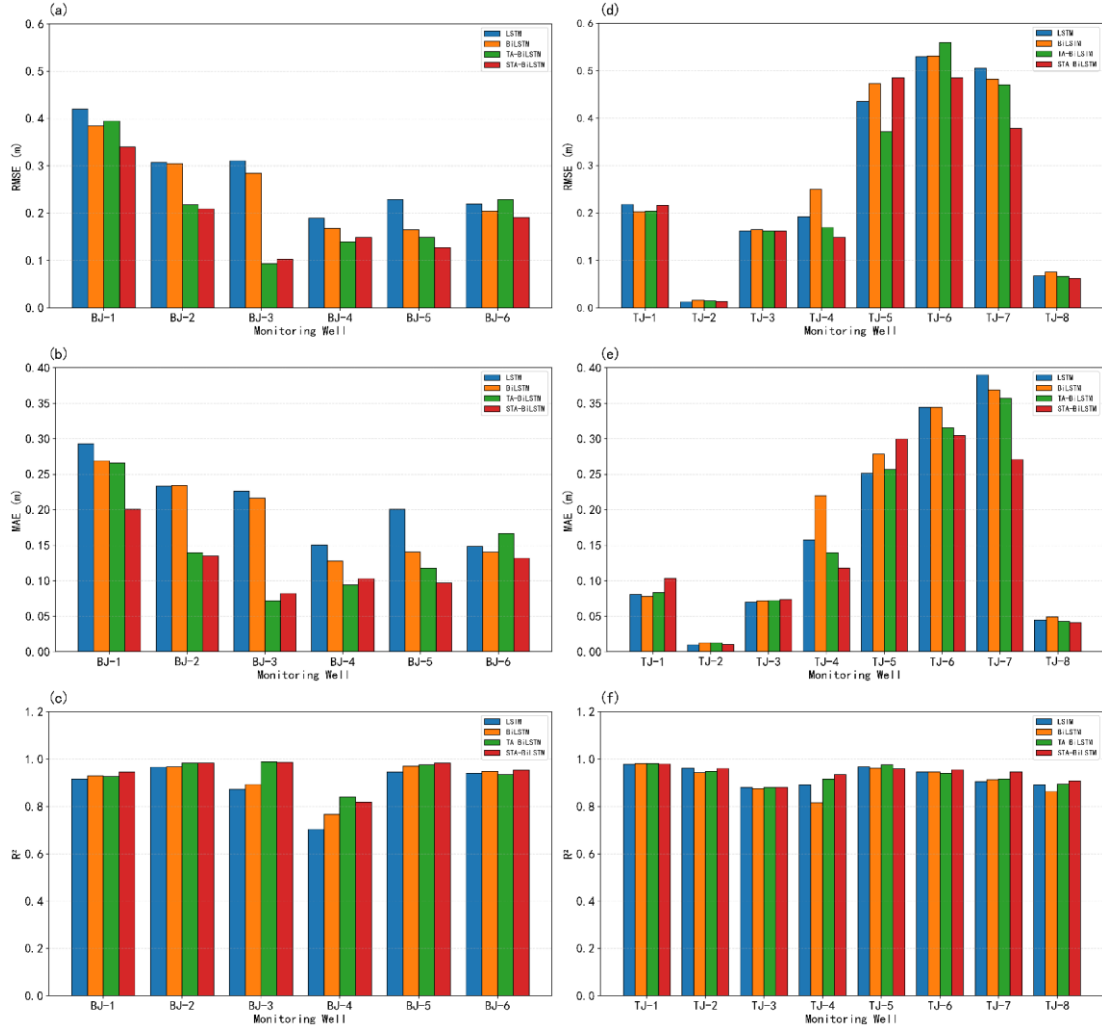


Figure 3: Comprehensive comparison of model performance metrics across 14 monitoring wells. (a-c): RMSE, MAE and R^2 comparison for Beijing rising zone wells. (d-f): RMSE, MAE and R^2 comparison for Tianjin declining zone wells.

5.3. Prediction Curves for Typical Sites

As shown in Figure 4(a-d), to visually demonstrate the superiority of the STA-BiLSTM model in capturing time-series details, two monitoring wells were selected from both Beijing and Tianjin to display and analyze the observed values versus the predicted values of each model on the test set.

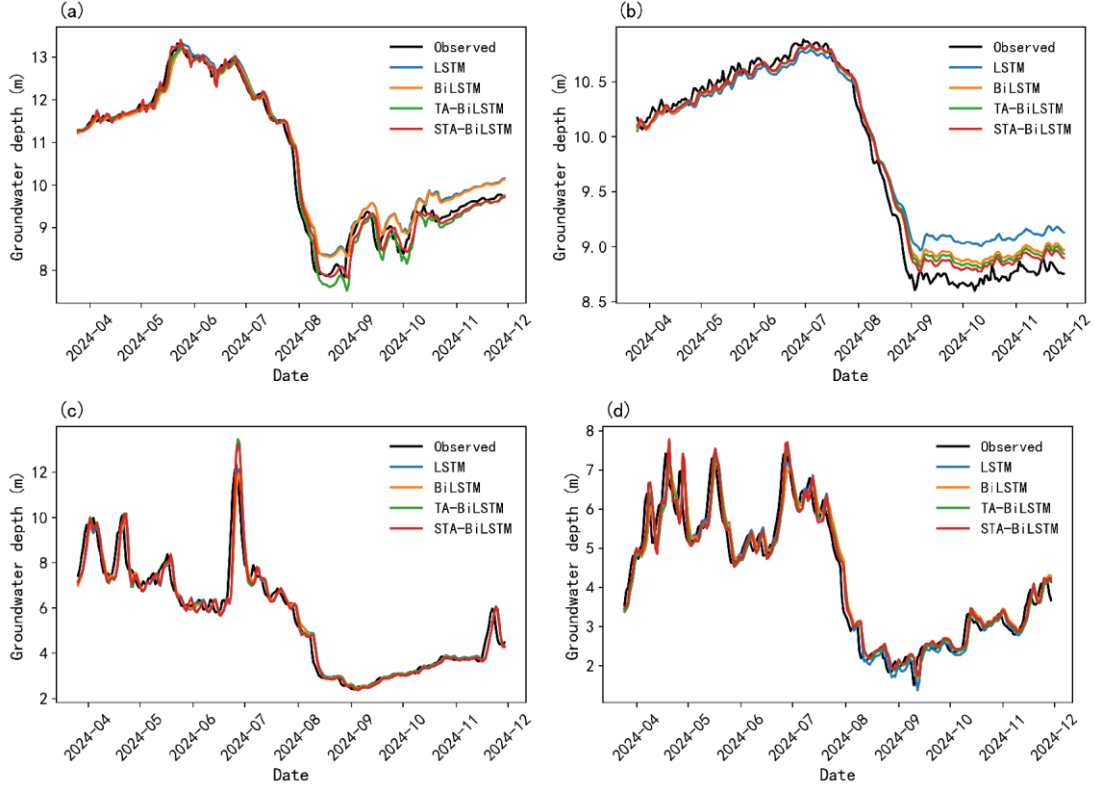


Figure 4: Observed vs predicted groundwater levels at representative wells. (a) BJ-2 well; (b) BJ-3 well; (c) TJ-6 well; (d) TJ-7 well.

At well BJ-3, groundwater depth exhibited a slow rise before August, followed by a sharp decline of approximately 1.5 m between August and September. LSTM and BiLSTM models displayed evident prediction inertia after this abrupt change, persistently overestimating depth. In contrast, the STA-BiLSTM model rapidly adjusted to align closely with the observed values thereafter. This demonstrates the advantage of spatiotemporal attention mechanisms: during abrupt shifts in a well's historical sequence, the model leverages synchronous or responsive information from spatial neighbors to quickly recalibrate the prediction baseline, thereby mitigating inertia errors inherent in single-sequence forecasting. In comparison, well BJ-2 showed a recovery trend following a similar decline, where all models performed better than at BJ-3. The prolonged low-level steady state after the abrupt change at BJ-3 hindered traditional LSTM from overcoming its memory inertia of historical sequences, whereas the evident dynamic recovery trend at BJ-2 facilitated timely adjustment of LSTM's learning direction, avoiding such prediction biases. At Tianjin sites like TJ-6 and TJ-7, STA-BiLSTM exhibits the highest overall fit to observed curves, particularly during fluctuating phases. This indicates that in regions with subdued trends, spatial information aids in capturing subtle fluctuations or regional synchronous changes that are imperceptible in single-well time series, thereby smoothing predictions and enhancing accuracy.

5.4. Model Robustness Analysis

To further assess the stability of model predictive performance, Figure 5(a-b) presents boxplots of RMSE values across all monitoring wells for each model in the Beijing and Tianjin regions. In both areas, the STA-BiLSTM model exhibits the lowest box position (median RMSE) and the shortest box length (smallest interquartile range), with medians significantly lower than those of other models. This indicates not only reduced average errors but also minimal performance

variability across sites, demonstrating strong robustness.

In contrast, LSTM and BiLSTM models show longer boxes with upper outliers, suggesting larger prediction errors and instability at sites with complex hydrogeological conditions. The TA-BiLSTM model improves stability over the former two, while STA-BiLSTM further compresses the error distribution range. This systematically validates the progressive enhancement from the baseline LSTM without attention, to the temporal attention mechanism, and finally to the spatiotemporal attention framework in addressing spatial heterogeneity and temporal non-stationarity.

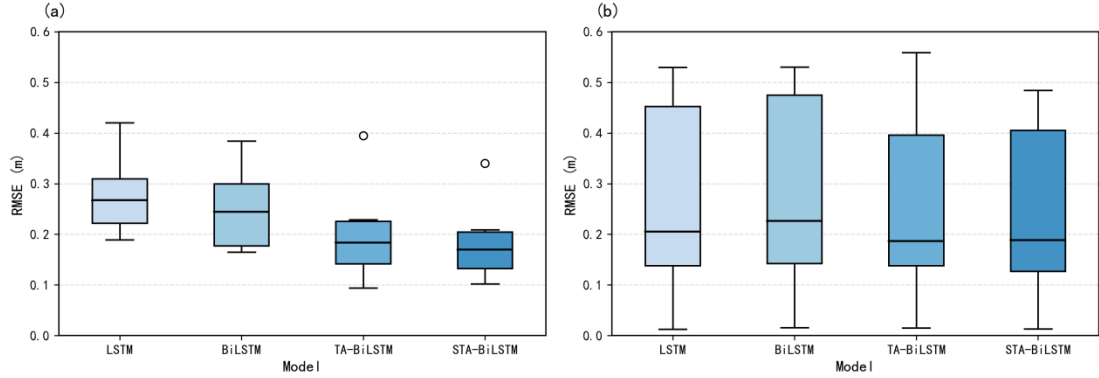


Figure 5: Boxplots of RMSE distribution across all monitoring wells in Beijing (a) and Tianjin (b) for different models.

6. Conclusions

The STA-BiLSTM model proposed in this study successfully integrates adaptive spatiotemporal feature extraction with Bayesian global optimization, effectively addressing the challenges of spatiotemporal dependency modeling and parameter optimization in groundwater level prediction. Empirical analysis in the Beijing–Tianjin region yields the following conclusions:

(1) Significant improvements in predictive accuracy. Compared with the baseline LSTM model, STA-BiLSTM reduces RMSE by 38.4% and MAE by 40.2% in the Beijing rising zone ($R^2 = 0.945$), and by 7.9% and 9.5% in the Tianjin declining zone ($R^2 = 0.940$). These results validate the model's superior capability in handling non-stationary sequences and complex fluctuation patterns, providing high-precision predictive support for fine-scale regional groundwater management.

(2) Spatiotemporal attention effectively resolves multidimensional dependencies.

The dual attention mechanisms effectively capture spatio-temporal correlations in multi-source monitoring data. The spatial attention module effectively aggregates information from neighboring monitoring wells through dynamic weight allocation, with the optimized statistically effective neighborhood radius exhibits potential correlations with regional hydrogeological conditions. The temporal attention module precisely focuses on key historical evolutionary moments. The synergy between the two modules substantially enhances prediction accuracy.

(3) Bayesian optimization enables adaptive model configuration. The algorithm automatically optimizes hyperparameters and structural parameters, with optimal spatial radii and attention orders varying across regions in accordance with local hydrogeological conditions. This adaptive mechanism ensures robust cross-regional applicability. Bayesian optimization achieves adaptive tuning of model structure and parameters.

In summary, the STA-BiLSTM model demonstrates high prediction accuracy and reveals underlying physical patterns through adaptive parameters, providing reliable technical support for groundwater management, ecological recharge assessment, and regional water resource

optimization.

Acknowledgements

This work was supported by the National Social Science Fund of China (NSSF) [Grant Number 23BZX103] (Project Title: Research on the Epistemic Opacity of Intelligent Machines).

References

- [1] Liu Y, Wu G, Ma B et al (2024) Revealing climatic and groundwater impacts on the spatiotemporal variations in vegetation coverage in marine sedimentary basins of the North China Plain, China. *Sci Rep* 14:10085. <https://doi.org/10.1038/s41598-024-60838-5>
- [2] Long D, Xu Y, Cui Y et al (2025) Unprecedented large-scale aquifer recovery through human intervention. *Nat Commun* 16:7296. <https://doi.org/10.1038/s41467-025-62719-5>
- [3] Su G, Xiong C, Zhang G, Wang Y, Shen Q, Chen X, An H, Qin L (2023) Coupled processes of groundwater dynamics and land subsidence in the context of active human intervention: a case in Tianjin, China. *Sci Total Environ* 903:166803. <https://doi.org/10.1016/j.scitotenv.2023.166803>
- [4] Zhang R, Wang T, Liu G, Shama A, Liu A, Wang X, Deng L, Mao W (2025) Ground subsidence evolution and groundwater variation in the middle route of South-to-North Water Diversion Project. *Sci Rep* 15:19220. <https://doi.org/10.1038/s41598-025-04307-7>
- [5] Long D, Yang W, Scanlon BR, Zhao J, Liu D, Burek P, Pan Y, You L, Wada Y (2020) South-to-North Water Diversion stabilizing Beijing's groundwater levels. *Nat Commun* 11:3665. <https://doi.org/10.1038/s41467-020-17428-6>
- [6] Wunsch A, Liesch T, Broda S (2021) Groundwater level forecasting with artificial neural networks: a comparison of long short-term memory (LSTM), convolutional neural networks (CNNs), and non-linear autoregressive networks with exogenous input (NARX). *Hydrol Earth Syst Sci* 25:1671–1687. <https://doi.org/10.5194/hess-25-1671-2021>
- [7] Thakur A, Chandel A, Shankar V (2025) Prediction of groundwater levels using a long short-term memory (LSTM) technique. *J Hydroinform* 27(1):51–68. <https://doi.org/10.2166/hydro.2024.239>
- [8] Zhang X, Dong F, Chen G, Dai Z (2023) Advance prediction of coastal groundwater levels with temporal convolutional and long short-term memory networks. *Hydrol Earth Syst Sci* 27:83–96. <https://doi.org/10.5194/hess-27-83-2023>
- [9] Liu J, Shen C, O'Donncha F, Song Y, Zhi W, Beck HE, Bindas T, Kraabel N, Lawson K (2025) From RNNs to Transformers: benchmarking deep learning architectures for hydrologic prediction. *Hydrol Earth Syst Sci* 29:6811–6828. <https://doi.org/10.5194/hess-29-6811-2025>
- [10] Nazari A, Jamshidi M, Roozbahani A, Golparvar B (2025) Groundwater level forecasting using empirical mode decomposition and wavelet-based long short-term memory (LSTM) neural networks. *Groundw Sustain Dev* 28:101397. <https://doi.org/10.1016/j.gsd.2024.101397>
- [11] Nourani V, Khodkar K, Gebremichael M (2022) Uncertainty assessment of LSTM-based groundwater level predictions. *Hydrol Sci J* 67(5):773–790. <https://doi.org/10.1080/02626667.2022.2046755>
- [12] Zhao Y, Yang L, Pan H, Li Y, Shao Y, Li J, 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. *J Environ Sci* 153:128–142. <https://doi.org/10.1016/j.jes.2024.03.052>
- [13] Chen Q, Jiang L, Qin H, Kontar RA (2025) Multi-agent collaborative Bayesian optimization via constrained Gaussian processes. *Technometrics* 67:32–45. <https://doi.org/10.1080/00401706.2024.2365732>
- [14] Liu X-D, Zhang H-Y, Rui X-P, Sun W (2025) Multi-factor LSTM groundwater level prediction model with attention mechanism. *Hydrology* 45(02):73–79. <https://doi.org/10.19797/j.cnki.1000-0852.20240161>(in Chinese)
- [15] Ding Y, Zhu Y, Feng J, Zhang P, Cheng Z (2020) Interpretable spatio-temporal attention LSTM model for flood forecasting. *Neurocomputing* 403:348–359. <https://doi.org/10.1016/j.neucom.2020.04.110>
- [16] Ouyang S, Xu C, Xu W, Zhou M, Zhang J, Zhang G, Pan Z (2025) Intelligent decoupling of hydrological effects in Han River cascade dam system: spatial heterogeneity mechanisms via an LSTM-attention-SHAP interpretable framework. *Hydrology* 12(8):217. <https://doi.org/10.3390/hydrology12080217>
- [17] Wang L, Jiang Z, Song L, Yu X, Yuan S, Zhang B (2024) A groundwater level spatiotemporal prediction model based on graph convolutional networks with a long short-term memory. *J Hydroinform* 26(11):2962–2979. <https://doi.org/10.2166/hydro.2024.226>
- [18] Zhao X, Dong S, Rao H, Ming W (2025) Water flow forecasting model based on bidirectional long- and short-term memory and attention mechanism. *Water* 17(14):2118. <https://doi.org/10.3390/w17142118>
- [19] Han Z, Li F, Zhao Y, Liu C (2025) Investigation into groundwater level prediction within a deep learning framework: incorporating the spatial dynamics of adjacent wells. *J Hydrol* 657:133097.

<https://doi.org/10.1016/j.jhydrol.2025.133097>

- [20] Shen C (2018) A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resources Research* 54:8558–8593. <https://doi.org/10.1029/2018WR022643>
- [21] Belete DM, Huchaiah MD (2022) Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. *Int J Comput Appl* 44:875–886. <https://doi.org/10.1080/1206212X.2021.1974663>
- [22] Luo G (2016) A review of automatic selection methods for machine learning algorithms and hyper-parameter values. *Netw Model Anal Health Inform Bioinform* 5:18. <https://doi.org/10.1007/s13721-016-0125-6>
- [23] Wu J, Chen XY, Zhang H, Xiong LD, Lei H, Deng SH (2019) Hyperparameter optimization for machine learning models based on Bayesian optimization. *Journal of Electronic Science and Technology* 17:26–40.
- [24] Yegin AE, Karcioglu AA (2025) Dam water levels prediction using advanced hybrid deep learning model based on Bayesian optimization approach. *Egypt Inform J* 31:100760. <https://doi.org/10.1016/j.eij.2025.100760>
- [25] Hochreiter S, Schmidhuber J (1997) Long Short-Term Memory. *Neural Computation* 9(8):1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- [26] Graves A, Schmidhuber J (2005) Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw* 18:602–610. <https://doi.org/10.1016/j.neunet.2005.06.042>
- [27] Siami-Namini S, Tavakoli N, Siami Namin A (2019) The performance of LSTM and BiLSTM in forecasting time series. 2019 IEEE International Conference on Big Data (Big Data), pp 3285–3292. <https://doi.org/10.1109/BigData47090.2019.9005997>
- [28] Vaswani A, Shazeer N, Parmar N, et al. (2017) Attention is all you need. *Advances in Neural Information Processing Systems* 30 (NIPS 2017), pp 5998–6008.
- [29] Shahriari B, Swersky K, Wang Z, Adams RP, de Freitas N (2016) Taking the human out of the loop: a review of Bayesian optimization. *Proc IEEE* 104:148–175. <https://doi.org/10.1109/JPROC.2015.2494218>
- [30] Chen C, Zhu X, Kang X, Zhou H (2021) A deep learning algorithm for groundwater level prediction based on spatial-temporal attention mechanism. 2021 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCCom/CyberSciTech), AB, Canada, 2021, pp 716–723. <https://doi.org/10.1109/DASC-PiCom-CBDCCom-CyberSciTech52372.2021.00119>
- [31] Cui J, Yang B (2018) Review of Bayesian optimization methods and applications. *Journal of Software Science* 29(10):3068–3090. <https://doi.org/10.13328/j.cnki.jos.005607>(in Chinese)
- [32] Patro, S Gopal & Sahu, Dr-Kishore Kumar. (2015) Normalization: A Preprocessing Stage. *IARJSET*. <http://doi.org/10.17148/IARJSET.2015.2305>
- [33] Franceschi L, Donini M, Perrone V, Klein A, Archambeau C, Seeger M, Pontil M, Frasconi P (2025) Hyperparameter optimization in machine learning. *Foundations and Trends in Machine Learning* 18(6):975–1109. <https://doi.org/10.1561/22000000088>