



## Research article

## Machine learning-based techniques for land subsidence simulation in an urban area

Jianxin Liu<sup>a,b</sup>, Wenxiang Liu<sup>a,c</sup>, Fabrice Blanchard Allechy<sup>d,e</sup>, Zhiwen Zheng<sup>f</sup>, Rong Liu<sup>a,b</sup>, Kouao Laurent Kouadio<sup>a,b,d,\*</sup><sup>a</sup> School of Geosciences and Info-physics, Central South University, Changsha, Hunan, 410083, China<sup>b</sup> Hunan Key Laboratory of Nonferrous Resources and Geological Hazards Exploration, Changsha, Hunan, 410083, China<sup>c</sup> Guangdong Geological Bureau, Guangzhou, Guangdong, 510700, China<sup>d</sup> UFR des Sciences de la Terre et des Ressources Minières, Université Félix Houphouët-Boigny, Abidjan, 22 BP 582 Abidjan 22, Côte d'Ivoire<sup>e</sup> Agricultural Research Centre for International Development (CIRAD), Montpellier, Occitanie, 34170, France<sup>f</sup> Guangdong Geological Environment Monitoring Station, Guangzhou, Guangdong, 510599, China

## ARTICLE INFO

Handling editor: Jason Michael Evans

## Keywords:

Land subsidence

Machine learning

Environmental risk assessment

Groundwater impact modeling

## ABSTRACT

Understanding and mitigating land subsidence (LS) is critical for sustainable urban planning and infrastructure management. We introduce a comprehensive analysis of LS forecasting utilizing two advanced machine learning models: the eXtreme Gradient Boosting Regressor (XGBR) and Long Short-Term Memory (LSTM). Our findings highlight groundwater level (GWL) and building concentration (BC) as pivotal factors influencing LS. Through the use of Taylor diagram, we demonstrate a strong correlation between both XGBR and LSTM models and the subsidence data, affirming their predictive accuracy. Notably, we applied delta-rate ( $\Delta r$ ) calculus to simulate a scenario with an 80% reduction in GWL and BC impact, revealing a potential substantial decrease in LS by 2040. This projection emphasizes the effectiveness of strategic urban and environmental policy interventions. The model performances, indicated by coefficients of determination  $R^2$  (0.90 for XGBR, 0.84 for LSTM), root-mean-squared error RMSE (0.37 for XGBR, 0.50 for LSTM), and mean-absolute-error MAE (0.34 for XGBR, 0.67 for LSTM), confirm their reliability. This research sets a precedent for incorporating dynamic environmental factors and adapting to real-time data in future studies. Our approach facilitates proactive LS management through data-driven strategies, offering valuable insights for policymakers and laying the foundation for sustainable urban development and resource management practices.

## 1. Introduction

Land subsidence (LS) is regarded as one of the most serious natural hazards that can occur unexpectedly, resulting in significant property damage such as building foundations, transit networks, underground pipelines, drainage systems, and other infrastructures (Chaussard et al., 2014; Pacheco-Martinez et al., 2013; Rahmati et al., 2019a; Van Niekerk and der Walt, 2006; Yin et al., 2016). It is a geohazard and visible indicator of land degradation caused by either natural or manmade factors such as poor land management, overexploitation of groundwater, and urban and agricultural development. The phenomenon mostly happens in many arid and semi-arid areas (Budhu and Adiyaman, 2010; Motagh et al., 2008). Moreover, it is frequently a source of concern since it

reduces an aquifer's storage capacity, resulting in geological breaches, fissures, damage to civil infrastructure, and increased flood risk. LS has emerged as a global threat, affecting numerous countries. Notable studies illustrating this include Wang et al. (2023) in China, Ebrahimy et al. (2020) in Iran, Corbeau et al. (2019) in Italy, Brown and Nicholls (2015) in Bangladesh, Chaussard et al. (2014) in Mexico and Galloway and Burbey (2011) in the United States of America. In recent decades, the prevalence of LS in some countries, like Iran, has skyrocketed (Motagh et al., 2008). Indiscriminate groundwater extraction for agricultural activities has been identified as a primary cause of LS in Iran (Foroughnia et al., 2019; Mohammady et al., 2019). Nonetheless, LS is a complex phenomenon influenced by a variety of factors beyond just groundwater extraction. Recognizing the widespread concern over LS,

\* Corresponding author. School of Geosciences and Info-physics, Central South University, Changsha, Hunan, 410083, China.

E-mail addresses: [ljx@csu.edu.cn](mailto:ljx@csu.edu.cn) (J. Liu), [liuwenxiang@csu.edu.cn](mailto:liuwenxiang@csu.edu.cn) (W. Liu), [Blanchard.fabrice.allechy@cirad.fr](mailto:Blanchard.fabrice.allechy@cirad.fr) (F.B. Allechy), [13539770009@163.com](mailto:13539770009@163.com) (Z. Zheng), [liurongkaoyan@csu.edu.cn](mailto:liurongkaoyan@csu.edu.cn) (R. Liu), [lkouao@csu.edu.cn](mailto:lkouao@csu.edu.cn), [etanoyau@gmail.com](mailto:etanoyau@gmail.com) (K.L. Kouadio).<https://doi.org/10.1016/j.jenvman.2024.120078>Received 6 December 2023; Received in revised form 4 January 2024; Accepted 6 January 2024  
0301-4797/© 2024 Elsevier Ltd. All rights reserved.

recent scientific endeavors from various corners of the world, encompassing the United States (Ellis et al., 2023), Japan (Nishi et al., 2023), China (Li et al., 2023; Shi et al., 2020a), Australia (Pan et al., 2022), France (Charpentier et al., 2022), Mexico (Fernández-Torres et al., 2022), Argentina (Solorza et al., 2022), Indonesia (Hayati et al., 2022), India (Raju et al., 2022), and Iran (Rahmati et al., 2019b; Shahbazi et al., 2022) have contributed significantly to our understanding of this global phenomenon. These studies, spanning diverse geographic contexts, collectively highlight LS as a pressing issue across different environments and urban settings.

This global perspective is exemplified in the case of Bangkok (Phien-wej et al., 2006). In this city, LS has presented a critical urban challenge for over 35 years, primarily driven by groundwater extraction. The persistence of this problem, despite efforts to mitigate it, underscores the multifaceted nature of LS, where human influences intersect with geological processes. Similarly, Abidin et al. (2011) document the varying rates of LS in Jakarta, Indonesia, attributing them to a combination of groundwater extraction, construction loads, and natural soil dynamics. This research highlights the complex and varied aspects of LS in rapidly developing urban areas and its ongoing impact on urban planning. Shirzaei et al. (2021) extend the scope of LS impact to coastal regions, identifying both natural and anthropogenic causes as significant contributors to relative sea-level rise and increased flooding hazards. Their review calls for the development of multi-objective predictive models, integrating physical and socio-economic factors, to accurately project coastal subsidence patterns. In China, as Xue et al. (2005) report, LS is primarily driven by excessive groundwater withdrawal, with further contributions from oil and warm groundwater extraction, and neotectonic movements. The study highlights the persistent and expanding nature of LS in China, revealing the complexities of geological and human-induced factors in exacerbating this environmental issue. Bagheri-Gavkosh et al. (2021) provide a global perspective by examining 290 LS cases, noting that a significant portion occurs in coastal and river deltaic regions, largely due to groundwater extraction. This review emphasizes the critical role of spaceborne monitoring techniques in understanding LS dynamics. Complementing this global view, Marfai and King (2007) explore LS in Semarang, Indonesia, using Digital Elevation Models (DEMs) and Geographic Information System (GIS) raster operations. Their study, focused on monitoring and predicting subsidence, predicts that by 2020 an area of 27.5 ha in Semarang will be 1.5–2.0 m below sea level, assuming a linear continuation of current subsidence rates without any mitigative action. This finding highlights the utility of DEMs and GIS in assessing and forecasting LS, particularly in urban areas facing significant infrastructural challenges. These investigations have repeatedly demonstrated that, due to the widespread distribution of LS and its potentially disastrous effects on the economy and environment, there is an urgent need for LS sensitivity zone assessments and identification of the leading causes of LS.

In China, the rapidly urbanizing regions, such as the Nansha district in Guangzhou, Guangdong Province, have caught the attention of researchers, urban planners, and policymakers. Initial studies into LS in the Guangdong province primarily revolved around natural causes, such as tectonic activities and sediment compaction. However, as urbanization increased, anthropogenic factors became more prevalent in discussions about subsidence. Urban construction, especially the development of skyscrapers and underground infrastructure, has also been noted to contribute to the pressure on the land, leading to subsidence. In the past decade, the occurrence of land-surface sinking has given rise to geologically induced hazards that pose a significant risk to the safety and well-being of urban populations. These risks include building cracks, ground fissures, underground pipe bursts, and bridge sinks. In instances of this nature, the implementation of rapid LS monitoring and precise simulation techniques can enhance the effectiveness of prevention and conservation endeavors (Du et al., 2021; Zhou et al., 2020). Semi-theoretical models and empirical models have

been utilized in LS simulation for a considerable period, as evidenced by their application in prior research (Guzy and Malinowska, 2020). Wolkersdorfer and Thiem (1999) employed a hydrogeological model to simulate LS phenomena in Germany, while Tang et al. (2008) made predictions regarding LS in Shanghai through the utilization of a grey model. The findings of Deng et al. (2017) indicate that conventional approaches have been inadequate in addressing these issues. LS is a multifaceted geological phenomenon resulting from the convergence of various variables, rendering it a non-linear issue (Li et al., 2021).

Recent statistics from the National Bureau of Statistics reveal that Guangdong Province's, China, urban landscape, particularly in the Pearl River Delta (PRD) - encompassing dynamic cities like Guangzhou, Hong-Kong, Macao, Foshan, Shenzhen, and Dongguan - is undergoing a remarkable transformation. As these cities expand, the PRD region faces a significant decline in sediment accumulation, compounded by the presence of multiple layers of soft soil with suboptimal mechanical properties. This scenario, coupled with intensive human engineering activities and natural geological processes, has led to severe LS issues across the area. Current estimates suggest that LS affects approximately 11,397 km<sup>2</sup> in the PRD, manifesting in visible damages such as cracked roads, collapsing houses, and tilting grounds, with the economic toll ranging from tens of thousands to millions of dollars. This emerging challenge not only jeopardizes the safety of these burgeoning megacities but also poses a critical threat to the sustainable economic development of the region. Given these circumstances, there is a pressing need for a comprehensive risk assessment to preemptively address this issue. Our study plays a pivotal role in mapping the extent and intensity of LS in the PRD basin, providing invaluable insights into its present impact and potential future consequences.

The advent of satellite technology, especially Interferometric Synthetic Aperture Radar (InSAR), has revolutionized the monitoring of LS. For instance, Hongdong et al. (2011) demonstrated the effectiveness of the Differential-InSAR (D-InSAR) technique in monitoring LS, particularly in Jiangsu province, China. Their findings establish a linear relationship between subsidence rates and groundwater factors and also highlight the precision and advantages of D-InSAR in large-area deformation monitoring. Cigna and Tapete (2022), Hayati et al. (2022), Strozzi et al. (2001), Wang et al. (2023), and many other researchers have harnessed this technology to gain more accurate and expansive readings of land movement in urban areas such as the Nansha district. With spatial resolutions ranging from 5 × 20 m for Sentinel 1 to 2 × 3.3 m for TerraSAR-X, remote sensing technology derived from InSAR provides a more efficient and less expensive method of obtaining information (Bai et al., 2016). Indeed, InSAR determines the satellite-Earth distance by measuring the phase difference between two or more images. Numerous studies, including those by Galloway and Burbey (2011), Golian et al. (2021), and Othman and Abotalib (2019), have effectively combined INSAR methods with machine learning (ML) techniques to enhance our understanding of LS (Deng et al., 2017; Rahmati et al., 2019a, 2019b; Wang et al., 2023). While the use of a single dependent variable is commonplace, the use of several independent variables in ML is much less common.

Nowadays, ML has gained prominence as a cutting-edge methodology for addressing nonlinear issues. It has emerged as a promising avenue for studying the simulation and prediction of LS (Ghorbani et al., 2022; Li et al., 2021). For instance, Zhu et al. (2015) employed a Deep Neural Network (DNN) in conjunction with a Genetic Algorithm (GA) to model the land surface of Beijing. The findings of the study indicate that the DNN-GA model has the potential to effectively replicate LS. The model demonstrated an average absolute inaccuracy of 32 mm when comparing the simulated values to the actual values. In addition, Zhou et al. (2019) employed the Gradient Boosted Decision Tree (GBT) methodology to assess the relative significance of several factors in the occurrence of LS within the eastern region of the Beijing Plain. The study conducted by the researchers found that the primary determinants influencing the rate of LS are the groundwater level and the thickness of

compressible layers. The model achieved a level of accuracy of 0.74. Moreover, Shi et al. (2020b) employed a Long Short-Term Memory (LSTM) model to simulate the temporal fluctuations in LS between the years 2011 and 2015. The study revealed that the LSTM model exhibited favorable performance in cases of little subsidence as indicated by an average root mean squared error of 10.85 mm/a.

Despite advancements in LS research, there remains a notable gap in the application of more sophisticated, data-driven methods capable of capturing the multifaceted and dynamic aspects of LS in urban environments. While studies have utilized ML techniques, such as DNN and GBT, in LS modeling (Zhu et al., 2015; Zhou et al., 2019), the exploration of more advanced ML methodologies like LSTM and eXtreme Gradient Boosting Regressor (XGBR) in urban LS forecasting is still in its nascent stages. Moreover, previous research has often focused on a limited set of variables or used single-model approaches, which may not fully encapsulate the intricacies of LS phenomena (Shi et al., 2020b). Our study seeks to fill this research void by leveraging both LSTM and XGBR models, thus providing a more holistic analysis that integrates a broader spectrum of variables and harnesses the collective strengths of these models. In this context, our research pioneers the use of LSTM and XGBR methodologies for LS modeling in urban areas, particularly focusing on future risk prevention. The primary objectives of this study are twofold: firstly, to estimate the rate of LS change utilizing data from InSAR, geology, and boreholes, thereby unraveling the key variables influencing LS; and secondly, to assess the efficacy of LSTM and XGBR models in generating susceptibility zoning maps for the years 2024, 2030, 2040, and 2060. We introduce robust modeling strategies incorporating the delta-rate ( $\Delta r$ ) concept, a novel approach in LS simulation. The dual application of LSTM and XGBR models addresses the limitations inherent in single-model frameworks when dealing with complex scenarios (Gelete, 2023). Furthermore, we conduct a Taylor analysis to evaluate the predictive performance of each model individually. By forecasting LS-prone areas, our study aims to mitigate the impacts of LS development, enabling the delineation of LS-risk zones, recommendation of effective control measures, and formulation of strategies for

disaster prevention, mitigation, and informed land use planning.

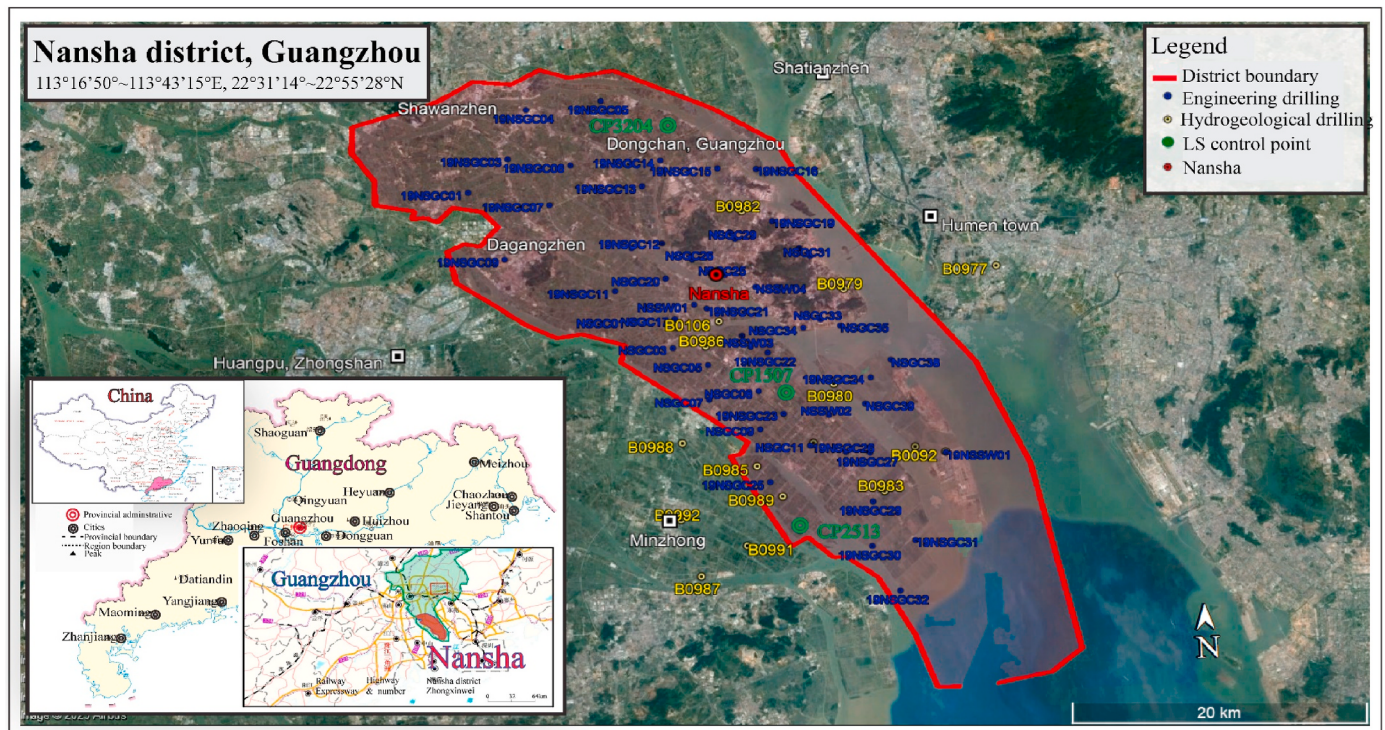
## 2. Study area and materials

In this section, we delve into the survey area and the materials used for our study. Focusing on the Nansha district, a rapidly urbanizing area within the dynamic Pearl River Delta (PRD) region, we explore various factors contributing to land subsidence (LS). Our materials encompass a variety of data that plays a crucial role in painting a comprehensive picture of the subsidence dynamics in this densely populated and industrially significant area.

### 2.1. Survey area

Nansha district is located at the southernmost tip of Guangzhou City along the west bank of the PRD Waterway and spans from 113°16'50" to 113°43'15" East longitude and 22°31'14" to 22°55'28" North latitude (Fig. 1). The district serves as the confluence point of the Xijiang, Beijiang, and Dongjiang rivers and covers a total area of approximately 803 square kilometers. The geological strata in Nansha are relatively complete, extending from the Sinian to the Quaternary period, except for the Silurian. This includes well-represented Paleogene and Quaternary formations, ranging from the oldest to the youngest. The Paleogene consists of the Xinzhuang formation, which occurs in some parts of the eastern and western parts of the Nansha District, with a small distribution area. The Quaternary strata are the most widely distributed and are spread over a large area in the delta plain, covering an area of 9814.1 km<sup>2</sup>, accounting for 72.9% of the total area of the PRD. The lithology is mainly characterized by soft soil deposition such as brown-red clay and sandy clay (Shiling Formation), yellowish gravel, medium coarse sand, spotted clay (Xiashi Formation), grey-white, brown-red gravel, gravel, and gravelly pebbles (Baini Formation).

In the Nansha area, hydrogeological conditions dictate the division of groundwater into two main aquifer layers: unconfined and confined. The unconfined aquifer, situated closer to the surface, is readily



replenished by rainfall and river water. However, its water is saline and not suitable for consumption. In contrast, the confined aquifer, trapped beneath less permeable layers like clay, is divided into two categories: the shallow confined aquifer at 25–100 m, and the deeper one ranging from 100 to 180 m.

This subterranean landscape, while rich in its complexity, faces challenges brought on by the broader trends of the PRD. Here, rapid urban development coupled with excessive groundwater extraction has become a primary contributor to LS. This phenomenon is especially pronounced in urban sectors of the PRD, including parts of Guangzhou, where accelerated urbanization has led to significant subsidence. The Nansha area, strategically situated within the PRD, displays a clear correlation between LS and the development of mega-infrastructure, especially near industrial zones. This suggests that urban expansion, not only in residential but also in industrial domains, plays a crucial role in exacerbating subsidence. Moreover, the impact of urbanization in Nansha extends beyond surface developments; it involves substantial groundwater extraction to support the growing population and industrial activities. This extraction leads to a decrease in groundwater levels, causing soil compaction and subsidence. Additionally, the weight of the constructed infrastructure, including high-rise buildings and industrial facilities, contributes to the ground pressure, further exacerbating the subsidence. Furthermore, geotechnical investigations, including engineering and hydrogeological drillings as shown in Fig. 1, reveal that LS in Nansha has progressed through four distinct stages. These stages align with the phases of urban development and infrastructure build-up:

- The beginning stage (2015–2016; ~125 mm): Initial signs of subsidence coinciding with early stages of rapid urban development and infrastructure construction.
- The evolving stage (2016–2018; ~170 mm): Increased subsidence rate corresponding with accelerated construction activities and heightened groundwater extraction.
- The expanding stage (2018–2020; ~222 mm): Further subsidence as urbanization spreads, with increased infrastructural load and continued groundwater use.
- The fast-developing stage (2020–2022; ~320 mm): A sharp rise in subsidence rate, likely due to the cumulative effects of sustained groundwater extraction, infrastructural development, and possibly other environmental factors such as soil composition and underlying geological features.

These stages highlight the interplay between human-induced changes (urbanization and groundwater extraction) and natural geological conditions, underscoring the complexity of LS in the Nansha district. The documentation suggests that without intervention and sustainable urban planning strategies, LS could continue to escalate, posing significant risks to the district’s infrastructure and overall environmental stability.

2.2. Materials

Our study utilizes an integrated dataset comprising Interferometric Synthetic Aperture Radar (InSAR) data for LS measurement, complemented by comprehensive geological and drilling data. The drilling dataset includes detailed engineering analyses of soil properties, such as compaction and soft soil characteristics, alongside groundwater level data essential for constructing accurate stratigraphic layer maps.

2.2.1. SAR image data

The Sentinel-1 satellite of the European Space Agency’s Copernicus program (GMES) was used. It provides continuous images (day, night, and various weather) and consists of two satellites, Sentinel-1A and Sentinel-1B. The sensors carried by the two satellites are synthetic aperture radars (SAR), which are active microwave remote sensing satellites. The shortest revisit period of a single Sentinel-1 satellite is 12

days. After the launch of Sentinel-1B, using the dual-satellite tandem flight method, the shortest revisit period is 6 days, which has good timeliness and reliability. Sentinel-1 carries a C-band synthetic aperture radar with a total of 213 scenes from June 15, 2015, to December 23, 2022. The parameters of Sentinel-1 are shown in Table 1. The SAR data is cross-checked with the real data collected via the control points every year (Fig. 2).

2.2.2. Geological data (Lith)

The geological data includes information on the geological systems that compose the region of Guangzhou, grouped into distinct compressible layers. The explicit lithology map was missing, and geological periods were used instead. Eight categories were retained to constitute our geological features (Lith), such as the Tertiary, the Cretaceous, the Jurassic, the Triassic, the Permian, the Carboniferous, the Devonian, and the Silurian systems (Fig. 3a). These data were obtained from the Guangdong Geological Bureau and do not change over time.

2.2.3. Filled soil thickness (FSt) and sand layer distributions (SLd)

Engineering and hydrogeological drilling are both types of drilling data collected in the Nansha district. They play a foundational role in ensuring that human activities, from construction to water extraction, are conducted safely, efficiently, and with minimal environmental impact. Engineering drilling, in particular, plays a pivotal role in assessing the subsurface ground conditions, which is fundamental to the design and construction of robust structures like buildings, bridges, tunnels, and dams. Data on filled soil thickness (FSt) and sand layer distributions (SLd) were meticulously gathered from these engineering drillings across different periods (Fig. 3b and c). Indeed, FSt and SLd are crucial because different soils can cause differential settlement if not adequately addressed. In addition, the distribution of soils determines the type of foundation needed for a structure; while sandy or gravelly soils might support shallow foundations, clay or silty soils might necessitate the use of deep foundations like pilings.

2.2.4. Building concentration (BC)

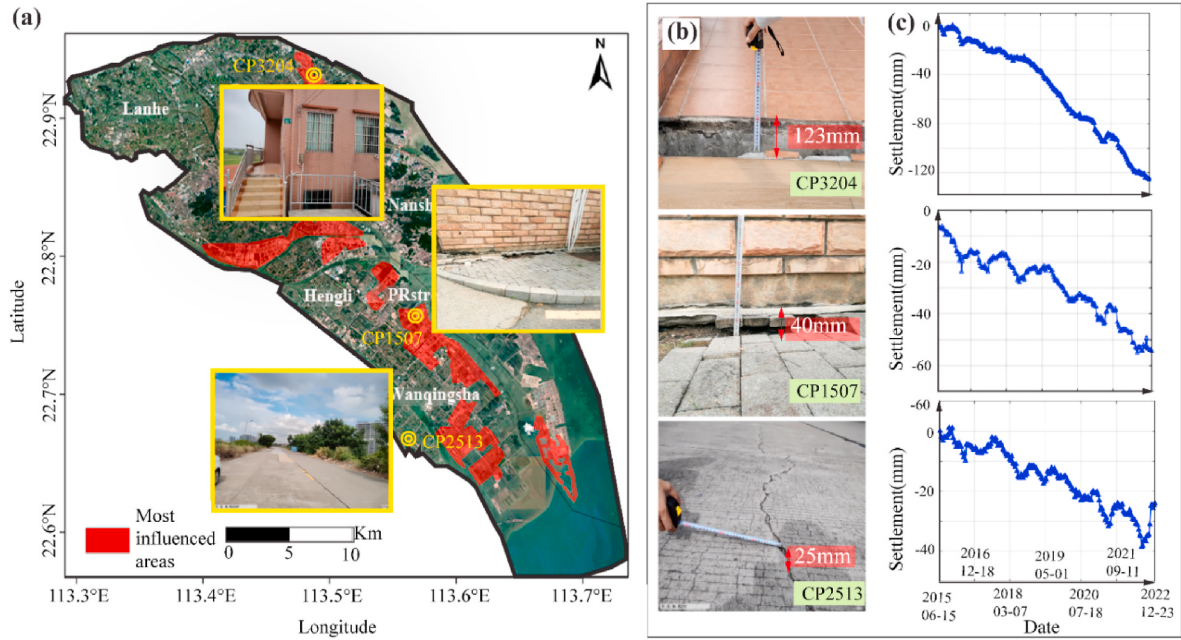
In the Nansha district, building concentrations (BC) are derived from comprehensive civil engineering data on structures. This is pivotal for our analysis, enabling a thorough exploration of the relationship between urban infrastructure density and LS in this densely populated area. The data, sourced from the Guangdong Geological Bureau, comprises a detailed classification of high-concentration zones of large buildings, covering the period from 2015 to 2022. The level of infrastructure is classified into three groups based on their height. The first class is composed of (C) low-rise buildings ranging from 3 to 10 stories, including detached homes, townhouses, small commercial buildings, residential apartments, offices, and mixed-use buildings. The second category (B) is composed of medium-rise buildings ranging from 10 to 40 stories, and the third class (A) is skyscrapers, supertall, and mega-tall buildings that exceed 40 stories (Fig. 3d).

2.2.5. Groundwater levels (GWL)

Hydrogeological drillings provide data on the presence, depth, quantity, and quality of groundwater. This information is vital for water

Table 1  
Parameters of Sentinel-1.

| Parameter                   | Sentinel-1A               | Sentinel-1B       |
|-----------------------------|---------------------------|-------------------|
| Beam modes                  | Interferometric Wide (IW) | Extra-Wide (EW)   |
| Wavelength                  | 5.6 cm                    | 5.6               |
| Band                        | C                         | C                 |
| Incidence angle             | 29°–46°                   | 29°–46°           |
| Repeat observations periods | 12                        | 6                 |
| Number of images            | 213                       | 213               |
| Data ranges                 | June 15, 2015             | December 23, 2022 |



**Fig. 2.** Samples of LS in Nansha district. The area marked in red shows the most severe subsidence based on control point results. They are mainly concentrated in Dagang Town, Wanqinsha, PR Street Villages, the southern dock of Longxue Street, and near Dongyoung Village. a) Examples of three control points. The latter were used to check the rate of discrepancies in subsidence each year. b) The diagram of three control points indicates some evolution of LS from June 2015 to December 2022. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

supply planning, especially in regions where groundwater is a primary source of potable water. The data collected are groundwater levels from a map resulting from the combination of all drillings collected in Guangdong Province. Fig. 3e shows the GWL of Nanshan extracted from the hydrogeological data spatial distributions of 2015 and 2018.

### 3. Methods

The methodologies are structured into a short overview of the SAR data processing step and the machine learning-based ensemble techniques.

#### 3.1. PS/SBAS-Interferometric Synthetic Aperture Radar (InSAR)

Due to atmospheric delay, orbital residuals, and decorrelation noise, InSAR technology can only measure surface deformation over two imaging periods. It can't measure deformation over time. This technology is also limited in terms of how it can be used. In recent years, to overcome the limitations, Tong et al. (2013) first proposed the Stacking technology, which reduces the impact of atmospheric delay by stacking multi-temporal differential interference phases. Subsequently, the mainstream time-series InSAR methods are invented and can be divided into two categories: the Permanent Scatterer Interferometry (PS-InSAR) based on a single image as the main image, and the interference point target analysis method developed on this basis. Another approach is the Small Baseline Subset (SBAS-InSAR) method, which utilizes a series of images with the main image as the focal point. This method emphasizes small baselines, employing techniques like Coherence Target (CT) and Temporal Coherence Point (TCP) for enhanced accuracy.

##### • PS-InSAR

The PS-InSAR largely overcame the influence of factors such as decoherence noise. The technique involves a series of specific processes, which can be outlined as follows:

- Stage 1: Select one image as the public reference image and perform its registration to ensure alignment with other images.
- Stage 2: Create a differential interferogram from any two aligned pictures. The phase value for any pair of photographs can be calculated using the following equation:

$$\varphi = \varphi_{flat} + \varphi_{to} + \varphi_{def} + \varphi_{am} + \varphi_n \quad (1)$$

where  $\varphi_{flat}$  is the earth's curvature,  $\varphi_{to}$  represents terrain undulations,  $\varphi_{def}$  signifies the deformation phase,  $\varphi_{am}$  represents atmospheric interference, and  $\varphi_n$  represents residual noise from the phase change.

- Stage 3: Extract permanent scatter points, invert surface displacement using phase information from neighboring points, and use the deformation model to remove atmosphere, elevation, and orbital errors.

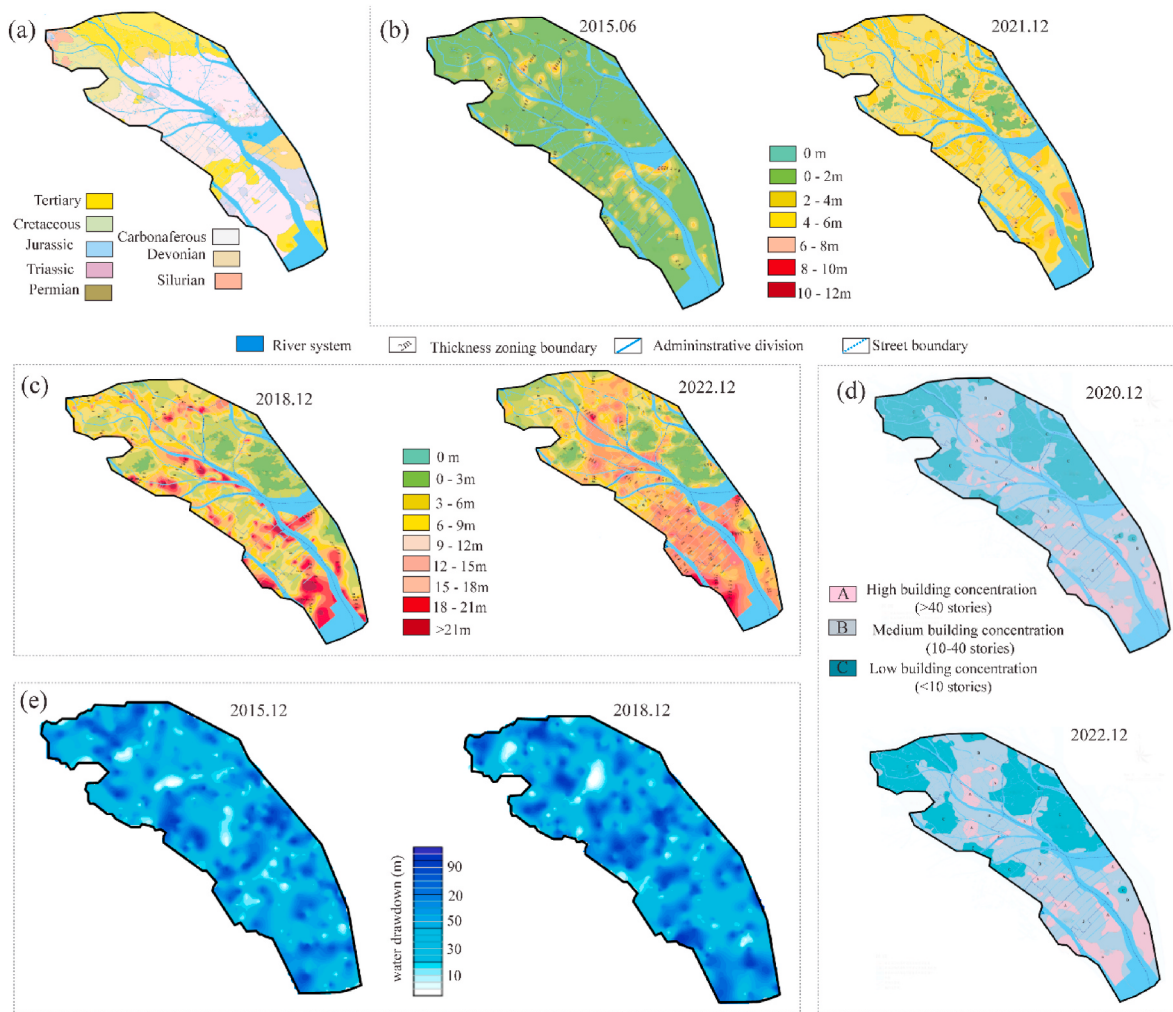
##### • SBAS-InSAR

Among the time series InSAR algorithms, the SBAS-InSAR (Tizzani et al., 2007) can better weaken the spatiotemporal decoherence caused by too long spatiotemporal baselines by selecting multiple main image interferences and can maximize the use of existing SAR data for high-speed accurate surface deformation inversion. The process of doing an SBAS-InSAR analysis involves the following steps:

- Stage 1: Create a differential interferogram between all picture pairings.
- Stage 2: Using least squares to compute the deformation phase as

$$\varphi = (A^T A)^{-1} A^T \Delta \varphi \quad (2)$$

in  $A$  – matrix, each row and column represent an interferogram and an image respectively.  $\Delta \varphi$  is a phase value combinatorial matrix on a differential interferogram. In the realm application case, it is the matrix  $A^T A$  is assumed a singular form, hence resulting in an unlimited number of solutions. The Singular Value Decomposition technique is employed to determine the solutions that minimize the sum of squared residuals



**Fig. 3.** Feature data. a) Geological, b) Filled Soil Thickness (FSt), c) Sand Layer Distribution (SLD), d) Building Concentration (BC) maps. The years of data collection for FSt and SLD are 2015 and 2018 for the first, and 2021 and 2022 for the second engineering work. e) Groundwater level (GWL) extracted from the whole Guangdong GWL map. The hydrogeological works were performed in 2015 and 2018.

for the parameters.

- Stage 3: involves the conversion of the deformation phase into a rate, followed by the determination of subsidence parameters for each period through the process of integration of these rates.

### 3.2. Learning machines

The learning machines are composed of eXtreme Gradient Boosting Regressor (XGBR) and Long Short-Term Memory (LSTM).

#### 3.2.1. Extreme Gradient Boosting Regressor (XGBR)

XGBR is a variant of a gradient-boosting algorithm used for regression tasks. It employs accurate approximations to construct an optimal prediction model (Friedman, 2001). Additionally, a notable benefit of XGBR is its efficiency in terms of computational speed. Additionally, it generates diverse training datasets through the process of random sampling. The idea behind the gradient boosting algorithm is to build a model in a stage-wise fashion by assembling a sequence of weak learners, typically decision trees, to create a strong learner that makes better predictions (Kouadio et al., 2022, 2023a). A simplified explanation of the mathematical operations involved in XGBR's objective function is given as

$$L(\varphi) = \sum l(y_i, \hat{y}_i) + \sum \Omega(f_i) \quad (3)$$

where  $L(\varphi)$  is the overall objective function,  $l(y_i, \hat{y}_i)$  is the loss function that measures the difference between the  $\hat{y}_i$  and the actual  $y_i$  values.  $\Omega(f_i)$  is the regularization term, which penalizes the complexity of the model (like the depth of trees). XGBR sequentially builds the model. At each step, it introduces a new tree  $f_k$  that predicts the residuals or errors of the previous model and adds it to the ensemble as

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_k(x_i) \quad (4)$$

where  $\hat{y}_i^{(t)}$  is the prediction at step  $t$  and  $x_i$  is the feature set for instance  $i$ .

#### 3.2.2. Long Short-Term Memory (LSTM)

The LSTM architecture is designed to address the constraints of conventional recurrent neural networks in capturing and learning long-term relationships (Kratzert et al., 2018). LSTM is capable of effectively governing and retaining information throughout temporal intervals. This characteristic renders the model very suitable for acquiring knowledge in long-term memory and capturing the impacts of dependencies (Koch and Schneider, 2022). In addition, LSTM comprises a memory cell responsible for storing information, along with three circular gates that regulate the flow of information within the LSTM. Refer

to [Appendix A1](#) for a complete mathematical explanation.

### 3.3. Data processing

The processing is divided into three distinct phases. Initially, we analyze SAR imagery to capture the primary data set. In the following, we focus on the extraction of groundwater level (GWL) data, a crucial step for understanding the influence of groundwater dynamics on LS. Lastly, we delve into the delta rate ( $\Delta r$ ) strategy applied with sophisticated machine learning (ML) algorithms to interpret and further analyze the data. This structured approach allows for a comprehensive analysis, starting from the nuanced interpretation of SAR images to advanced machine processing techniques, ensuring thorough data extraction and insightful analysis.

#### 3.3.1. SAR procedure

Two hundred twenty-five (213) scenes were used as SAR images collected from Sentinel 1- satellite from June 2015 to December 2022. PS-InSAR and SBAS-InSAR technologies were used to execute time-series InSAR surface deformation monitoring and processing operations, respectively. The PS-InSAR was used to process 144 scenes of photos obtained between June 2015–November 2019, with the image taken on December 2018, acting as the master image. Four steps summarize the procedure of SAR image processing.

- Data registration accuracy at any two separate time times met standards of within 1/8 cell by performing image registration for 144 sets of photos.
- The second step involved creating a differential interferogram using time-series interferometry on the registered images.
- The third step was the use of the amplitude difference technique to obtain PS points.
- Fourth, we used the deformation model to produce surface displacement data for the research area based on the information on permanent scatterer locations, minimizing errors associated with elevation, atmosphere, and orbit.

For the remaining 66 images, land surface deformation was tracked using SBAS-InSAR's time series captured between January 2019 and December 2022. Five hundred meters were used as the spatial baseline and 180 days as the temporal baseline during processing. A total of 45 pairs of images with minor baseline interference were established. The study area's land surface displacement data was acquired using the least squares approach and singular value decomposition.

Using PS-InSAR and SBAS-InSAR monitoring findings, the cumulative subsidence data from June 2015 -to November 2019 was used as benchmarks for the deformation sequence to implement time-series fusion for the cumulative deformation variables of June 2015 to December 2022.

#### 3.3.2. GWL extraction procedure

An aquifer map is drawn from hydrogeological drillings performed in the whole of Guangdong Province between June 2015 and December 2018. The map is built from the information collected in each borehole derived from the two aquifers (confined and unconfined). Thus, the hydrogeological data was analyzed using Golden Software Surfer 25.1 and then rasterized using ArcGIS Pro 1.2. The Nanshan GWL map is made by interpolating all the GWL data associated with each borehole from the province map (e.g., [Fig. 3e](#)). The semi-variogram ordinary Kriging interpolation model is used to make the map fit automatically. Next, the GWL contour lines of the two aquifers were created and transformed into a groundwater-level raster image.

#### 3.3.3. Delta-rate ( $\Delta r$ ) strategy

Data interpolation for time-series forecasting is necessary with the Nansha dataset. Indeed, all variables that compose the predictor are not

sampled according to the period from 2015 to 2022. For instance, the drilling data are a sampling of 2015 and 2018 ([Fig. 3](#)), whereas the soil-filled and sand layer engineering works were conducted in 2015 and 2021, and 2018 and 2022 ([Fig. 3b](#) and [c](#)), respectively. To fit all variables at the same time scale (i.e., from 2015 to 2022), we used the delta rate ( $\Delta r$ ) computation and applied it to all the data except the categorical feature Lith, which is expected to not vary across the time as

$$\Delta r (ts) = \frac{1}{ts} \left( \frac{D_{obs}(time\ 2) - D_{obs}(time\ 1)}{time\ 2 - time\ 1} \right) \quad (5)$$

where  $D_{obs}(time\ 2)$  and  $D_{obs}(time\ 1)$  are the data collected on dates 2 and 1 respectively.  $ts$  is a flexible timestamp and can be converted into months, weeks, days, etc. For instance, the  $\Delta r$  of FSt collected in 2017 and 2021 can be used to get the cumulative value of 2018, 2019, and 2020 as

$$D_{obs}(t+1) = D_{obs}(t) + \Delta r \quad (6)$$

where  $D_{obs}(t+1)$  and  $D_{obs}(t)$  are the observed data, we want to estimate at  $t+1$  from  $t$ -time respectively. This assumes that  $\Delta r$  is constant at all times by default. Finally, the predictors from 2015 to 2022 were constructed with  $60\text{ m} \times 60\text{ m}$  pixels from 2015 to 2022. To ensure that we have enough data from time-series calibration and validation, data were weekly sampled (timestamp) using the same approach as Eqs. (5) and (6) applied to two consecutive year-scale data. For instance, the  $\Delta r$  of 2021–2022 was divided into 53 weeks ( $ts = 53$ ), where  $\Delta r$  is constantly applied to each decomposed pixel that composes the LS features map. Overall, 395 weekly timestamps were created from June 2015 to December 2022. Thus, the average LS progression for each week should be used for the forecasting. Indeed, knowing the  $\Delta r$  in advance for each feature should be easily utilized for preventing LS forecasting. Thus, if the detected feature is an influential factor for LS, the application of reduction rate  $\tau$  into each  $\Delta r$  value should create a new feature value,  $\Delta r_{prev}$  utilized for retraining and re-evaluating the new prevention forecasting as

$$\Delta r_{prev} = \tau \times \overline{\Delta r} \quad (7)$$

where  $\Delta r_{prev}$  is the new delta-rate computed for prevention;  $\tau$  is the factor reduction for LS preventing the risk, and  $\overline{\Delta r}$  is a delta-rate averaged value of the time. Note that the use of averaged  $\overline{\Delta r}$  reduces the computation times applied to each pixel that constitutes the feature map. Indeed, the generated LS of a single map data for  $60\text{ m} \times 60\text{ m}$  pixels yields 274,967 samples which excessively consumes computation resources for training and validating multiple-variate times-series data.

However, if the influential factor is a categorical feature (for instance, Lith), the adjusted value using the down-weighting in Eq. (7) might not necessarily reduce the importance of a feature. We, therefore, propose two distinct techniques applied to LSTM and XGBR to work around these issues.

In LSTM, we implement the Lasso regression (L1 regularization) directly within an LSTM network to reduce the influence of a specific feature by adding a penalty equivalent to the absolute value of the magnitude of coefficients to the loss function  $\mathcal{L}$  as

$$\mathcal{L}_{new} = \mathcal{L}_{original} + \tau \lambda \sum |w| \quad (8)$$

Here,  $\lambda$  is the regularization strength ( $\lambda$  is recommended to be as small as possible; e.g.,  $\lambda = 10^{-6}$ ), and  $\sum |w|$  is the sum of the absolute values of the weights in the networks. The new Loss  $\mathcal{L}_{new}$  can lead to some coefficients being shrunk to zero, effectively reducing the influence of corresponding features. It is easier to implement since most deep learning frameworks, like TensorFlow ([Martín et al., 2015](#)) and PyTorch ([Paszke et al., 2019](#)) allow to add L1 regularization to the layers of a neural network, including LSTM.

In XGBR, the depth of the tree can be controlled to limit the model's complexity, indirectly affecting feature importance. Indeed, the depth of

a tree determines how many splits it makes before reaching a prediction. A shallower tree (with less depth) will generally have less capacity to fit complex patterns and therefore might rely less on features that contribute to these complex patterns. The new specification for the maximum depth of the tree ( $\max D_{new}$ ) for risk prevention at the  $\tau$ -factor threshold is expressed as

$$\max D_{new} = \tau \max D_{original} \quad (9)$$

### 3.3.4. Calibration and validation

In the context of utilizing black box models for modeling purposes, it is imperative to normalize both the input and output data. This normalization process ensures that all variables are brought into a consistent range before being input into the models. This approach will guarantee equitable consideration of all data, mitigate the influence of dimensionality, and prevent the overshadowing of data with tiny values by those in the higher numerical range (Nourani et al., 2019; Rajaei et al., 2019). Furthermore, the process of data normalization serves to streamline numerical computations within the model, thereby enhancing the precision of the modeling outcomes and diminishing the duration needed to ascertain the local or global minimum. In the present study, the process of normalizing the data was executed by employing Eq. (10), which effectively rescaled the values within the range of 0–1 as

$$X_n = (X_i - X_{min}) / (X_{max} - X_{min}) \quad (10)$$

where  $X_n$ ,  $X_{max}$ ,  $X_i$ , and  $X_{min}$  are used to denote the normalized, maximum, actual, and minimum values of the dataset, respectively. It is imperative to evaluate the precision of LS process simulation models throughout both the calibration and validation stages.

XGBR and LSTM are known for their effectiveness in modeling complex, non-linear patterns in time-series data. However, forecasting tasks, especially for XGBR, necessitate the first transformation of the time series dataset into a supervised learning issue. In addition, the utilization of the walk-forward validation technique is necessary for assessing the model. This is because employing k-fold cross-validation for model evaluation would yield overly optimistic results. The splits of the data in k-fold cross-validation are done in a way that is similar to randomness. This makes sure that the training and test/validation datasets are chosen from any part of the data that hasn't been used for training. To comprehend the significant LS predicting influential factors, the dataset was partitioned into two subsets: 70% and 30% for training and validation, respectively. However, when it comes to LS time-series data forecasting, this k-fold cross-validation may result in an excessive level of confidence in the model's performance. This is because time-series data typically changes over time. For this reason, walk-forward validation usually involves training a model on a small part of the dataset and then checking how well it did in a later period that comes after the training set. In this study, the model is trained using historical  $\Delta r$ -data and subsequently evaluated using the most recent subset of data. By eliminating the optimistic bias, a more accurate assessment of performance can be obtained. Regarding the utilization of training, validation, and test datasets, it is imperative to note that the adjustment of model parameters will be only conducted using the training and validation datasets. Moreover, the evaluation of the model's performance will be carried out by benchmarking against a test dataset that encompasses data that is temporally more advanced. In the walk-forward approach, we train on a small portion of the data and then test at a later time. Then we move on and repeat the procedure. As a result, we have many out-of-sample times and may combine the results over periods. The optimal configuration of hyperparameters for the models was determined through a process of trial and error, as illustrated in Table 2.

### 3.3.5. Model evaluation criteria

As suggested by Nourani et al. (2019), it is best to use at least one

**Table 2**

Hyperparameters for XGBR and LSTM.

| Model | Tuning parameter models                                                                                   |
|-------|-----------------------------------------------------------------------------------------------------------|
| XGBR  | n-estimators = 1000, maximum depth = 27, learning rate = 0.1, loss = squared error.                       |
| LSTM  | Number of hidden units = 200, max epoch = 100; optimization = Adamax; dropout = 0.4; learning rate = 0.01 |

statistical error metric and one goodness-of-fit measure when judging how well the models can predict the future. In this study, the accuracy of the proposed LSTM and XGBR was evaluated using mean absolute error (MAE), root-mean-square error (RMSE), and coefficient of determination ( $R^2$ ).

The MAE quantifies the discrepancy between expected and observed values, specifically by disregarding the impact of negative values given as

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \text{ with } 0 \leq MAE < \infty \quad (11)$$

where  $y$  is the observed data and  $\hat{y}$  is the predicted outcomes of the model of  $n$ -samples. A low MAE is indicative of a high level of accuracy in the predictions made by the model.

The RMSE measures the error of a model in predicting quantitative and runs from 0 to positive infinity. An ideal model would yield a root RMSE value of 0. It is expressed as

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \text{ with } 0 \leq RMSE < \infty \quad (12)$$

The coefficient of determination ( $R^2$ ) quantifies the extent of correlation or collinearity between the anticipated values and the actual values.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \text{ with } 0 \leq R^2 < 1 \quad (13)$$

where  $\bar{y}$  is the mean of the observed data. The range of the metric is from 0 to 1, where higher values signify a stronger alignment between the model and the observed data.

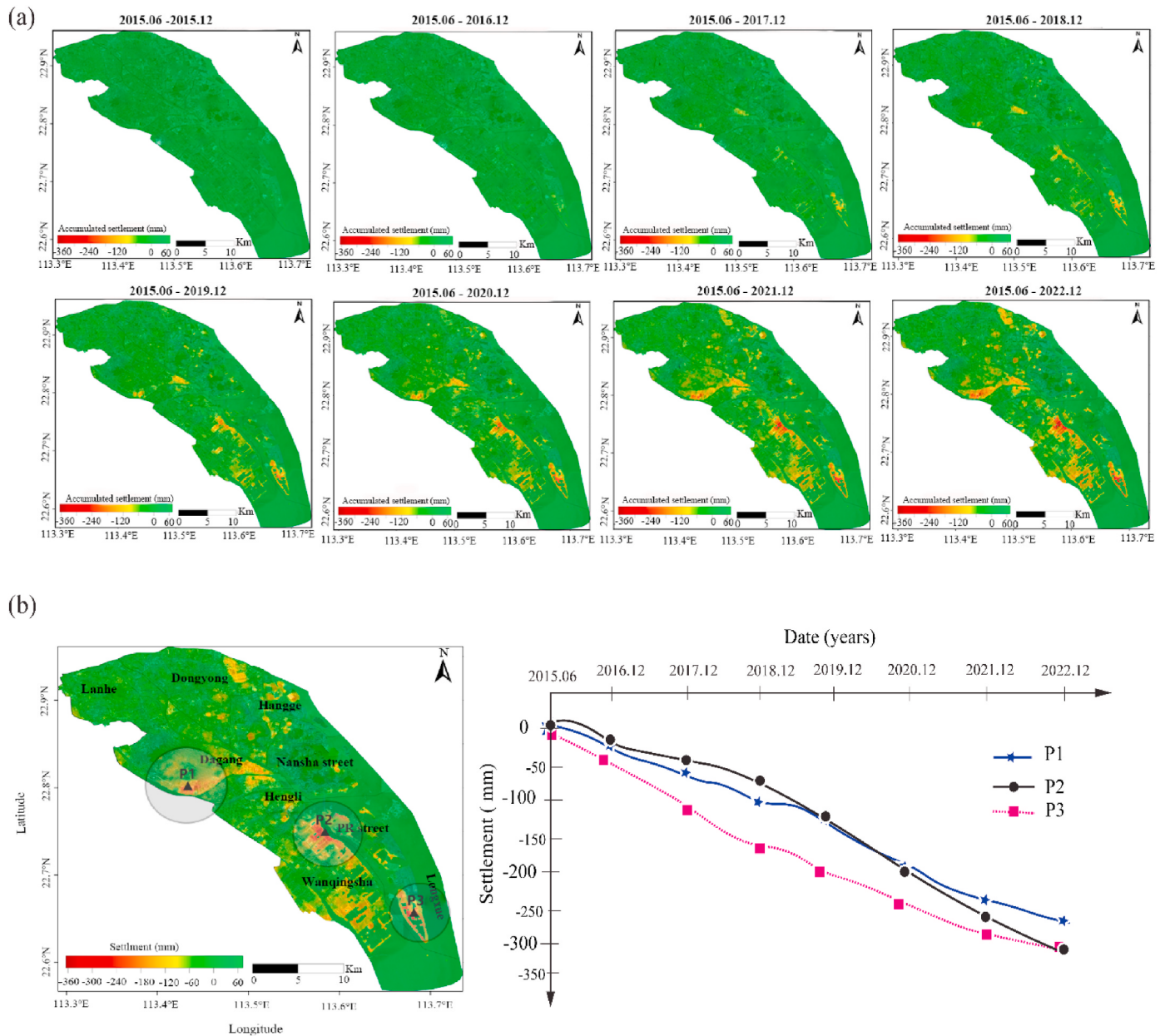
## 4. Results

This section provides a thorough analysis of land subsidence (LS), its forecasting, and risk prevention strategies in Nansha District. We begin with an analysis of LS's spatial and temporal patterns from 2015 to 2022, followed by a focused study of single-point subsidence evolution during the same period. Further, we investigate the factors contributing to LS, simulate the subsidence process, and evaluate the performance of our analytical methods. A Taylor Diagram visually illustrates our model's accuracy. Lastly, we propose strategies for LS prevention, offering comprehensive insights and contributing significantly to environmental and geotechnical research.

### 4.1. SAR

#### 4.1.1. Analysis of spatial and temporal patterns of subsidence in Nansha District from 2015 to 2022

The Nansha project calculated annual deformation findings (Fig. 4a). This scenario can be studied in seven categories. Thus, from 2015.06 to 2015.12, Nansha has no deformation signal and a maximum cumulative settlement of  $-21\text{mm}$ . The land and buildings were stable. Major deformations around many Nansha district streets occurred during 2015–2016, with the largest cumulative settlement reaching  $-64\text{mm}$ .



**Fig. 4.** InSAR maps. a) Cumulated LS map from 2015 to 2022. The map is rasterized and set input from 60mx60m pixels. b) LS control points (P1, P2, P3) evolution with their temporal evolution.

The southeast side showed the highest deformation from 2015.06 to 2017.12, with a maximum cumulative settlement of  $-128\text{mm}$ . From 2015 to 2018, the maximum cumulative settlement in Nansha District was  $-176\text{mm}$ , and from 2015 to 2019, it was  $-225\text{mm}$ . Moreover, the deformation on the west side was noticeable from 2015.06 to 2020.12, with a maximum total settlement of  $-263\text{mm}$ . However, the large-area deformation signals were found, in numerous communities from 2015.06 to 2021.12 with the highest cumulative settlement reaching  $-320\text{mm}$ . Finally, the greatest cumulative settlement was  $-364\text{mm}$  from 2015.06 to 2022.12.

#### 4.1.2. Analysis of single-point subsidence evolution in Nansha from 2015.06 to 2022.12

The study examined subsidence evolution in Nansha at numerous representative deformation locations in areas with considerable cumulative deformation. The researchers then created time series deformation graphs for each control point.

Fig. 4b shows the deformation site distribution of control points P1, P2, and P3, with their temporal evolution. Point P1's deformation rate remains stable over time. From 2015.06 to 2016.12, deformation in this area was minimal, then it began subsiding. Gradual stability in 2022.12 led to a maximum settlement of  $-254\text{ mm}$ . Over time, P2 tends to approach a condition of uniformity. The LS continues to evolve until the conclusion of 2022.12, displaying no discernible pattern of stabilization. Moreover, the maximum cumulative settlement attains a value of  $-313\text{ mm}$ . Besides, throughout the observed period from 2015.12 to 2016.12, there is a discernible, albeit modest, positive trajectory evident in the data points corresponding to P3. Subsequently, the object resumed its descent, gradually assuming a state of deformation that can be described as predominantly uniform. The settlement sum continued to increase without displaying any signs of stabilization until 2022.12. Moreover, the highest cumulative settlement reached a value in the vicinity of  $-350\text{ mm}$ .

## 4.2. Learning machines

### 4.2.1. Factor analysis

Fig. 5a illustrates the feature importance as derived from both the XGBR and Random Forest (RF) models (Ho, 1995). The inclusion of the RF model, known for its robustness and interpretability, serves as a comparative benchmark to validate and corroborate the influential factors identified by the XGBR model (Kouadio et al., 2023b). At a glance, GWL constitutes the most important feature, with the highest score suggesting that it has the greatest impact on the model's predictions, followed by BC. Besides, the feature contribution of RF differs from the XGBR model. Indeed, the building concentration BC represents the most significant feature, whereas the GWL comes in as the second most important feature. However, both models demonstrated that GWL and BC are the most influential factors and accumulate around 51% of importance. RF confirms the importance of GWL and BC as the main causes of the LS effect in the Nansha district.

### 4.2.2. LS simulation

#### • Forecast with delta rate ( $\Delta r$ )

Fig. 5b shows an LS predictive model in the Nansha district using the

averaged  $\overline{\Delta r}$ . It also assesses the impact of preventive measures. The actual measured subsidence rates up to the point labeled "Train End Date," which refers to the complete model training. Thus, from  $\overline{\Delta r}$ , the figure shows two scenarios for each model: a forecast without preventive measures and a forecast with preventive measures (lines with circle markers). The preventive measures involve reducing the GWL and BC by 80%. This reflects an attempt to mitigate subsidence by addressing its primary causes: groundwater extraction and the weight of dense urban infrastructure.

#### • Model confidence

The shaded areas around the forecast lines represent the 70–95% confidence intervals. This means there is a 70–95% probability that the actual subsidence rate will fall within these bounds. The confidence intervals widen over time, indicating increasing uncertainty in the predictions as we move further from the Train-End-Date. The actual subsidence rate appears to have been increasing, reaching closer to  $-45$  mm/year at the Train-End-Date. Moreover, the predictions without preventive measures continue this trend, with subsidence rates expected to worsen over time. The forecasts with prevention, however, show a marked improvement, flattening out the subsidence rate's downward trajectory. This suggests that reducing GWL and BC could significantly

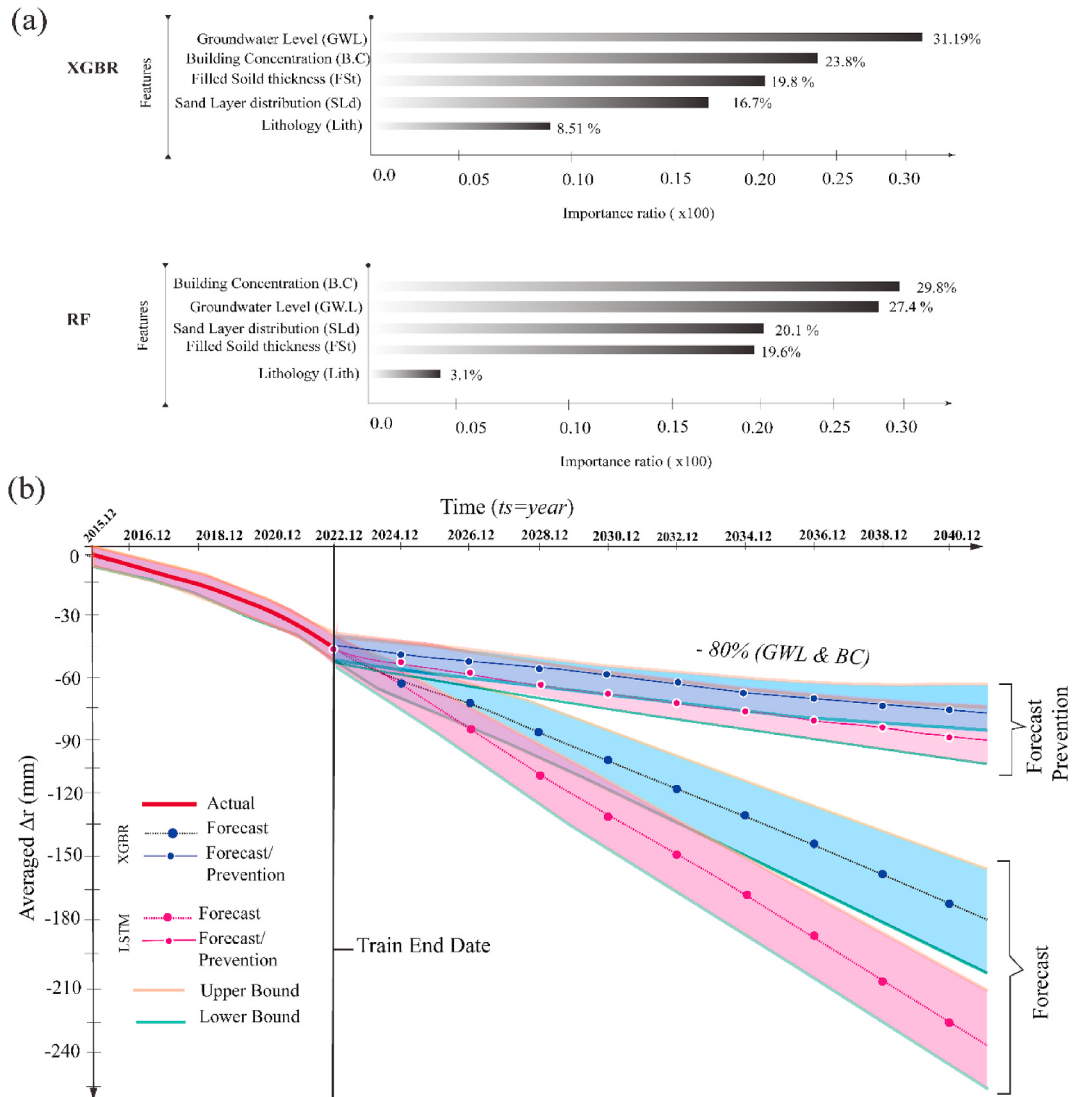


Fig. 5. Feature contribution during LS simulation. a) Feature importance evaluated using XGBR and Random Forest (RF). b) Forecast versus Actual using  $\Delta r$  strategy.

mitigate the rate of LS. Furthermore, there may be slight differences in the predictions of the XGBR and LSTM models. Indeed, XGBR might capture abrupt changes better, while LSTM might be better at capturing long-term dependencies.

#### • Predicted $\widehat{\Delta r}$ applied to LS data

Fig. 6a shows the XGBR and LSTM simulation thin 2060.12 using  $\widehat{\Delta r}$  applied to each pixel that composes the cumulated LS from the Training-End-Date (2022.12). Here, XGBR shows a relatively moderate amount of subsidence predicted from 2015 to 2024 with most areas less than 250 mm ( $\overline{\Delta r} = -65\text{mm}$ ). Some hotspots exceed 250 mm. The LS pattern from cumulative 2015–2030 becomes more pronounced, with wider areas reaching beyond the 250 mm mark ( $\overline{\Delta r} = -105\text{mm}$ ). The red areas from 2015 to 2040, indicate severe subsidence (exceeding 350 mm  $\sim \overline{\Delta r} = -173\text{mm}$ ). They are more extensive, suggesting a progressive worsening of the subsidence problem. This severity continues to grow, with several areas showing a predicted settlement of more than 450 mm from the cumulative at 2060 ( $\overline{\Delta r} = -254\text{mm}$ ). The scale for settlement changes with each consecutive map, indicating the predictive model expects a gradual increase in the amount of subsidence as time goes on. In Fig. 6b, using LSTM, the LS of 2015–2024 is distributed similarly to the XGBR model, but the areas with the most significant subsidence seem slightly less extensive. Moreover, a similar trend is observed with the XGBR model. However, the red areas appear slightly less intense in 2030 with  $\overline{\Delta r} = -137\text{mm}$ . From 2015 to 2040, the trend of increasing subsidence continues but seems slightly reduced in terms of intensity compared to the XGBR model. Like the XGBR model, the amount of subsidence appears to be increasing in 2060 ( $\overline{\Delta r} = -321\text{mm}$ ) but the extent and intensity of the most affected areas are less than what is predicted by the XGBR model.

Table 3 shows some of the main standout points of the two models, XGBR and LSTM. Furthermore, it is important to note that our data were interpolated using the  $\Delta r$  strategy (section 3.3.3) to fit LS week-time data; therefore, XGBR and LSTM predictions are subject to some uncertainties, and the true utility of these models would also depend on their validation against observed subsidence data.

**Table 3**

Key comparison between XGBR and LSTM models for LS simulation.

| Key points             | Scientific observations                                                                                                                                                                                                                                                                          |
|------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Model agreement        | XGBR and LSTM seem to agree on the general pattern of subsidence, with certain areas consistently showing as hotspots of severe subsidence across all future time frames.                                                                                                                        |
| Intensity differences  | LSTM model generally predicts more severe subsidence than the XGBR model. This could be due to differences in how the models weigh the input variables or the mathematical structure of the models themselves.                                                                                   |
| Progressive worsening  | XGBR and LSTM show a worsening trend over time, suggesting that whatever factors are contributing to the subsidence (such as groundwater depletion, natural compaction, or other geotechnical factors) are expected to continue or intensify.                                                    |
| Practical implications | Depending on the confidence in each model, planners and engineers might consider the LSTM predictions as a "worst-case" scenario and the XGBR as a "best-case" or less severe scenario. This could help in risk management and in setting more conservative or aggressive mitigation strategies. |

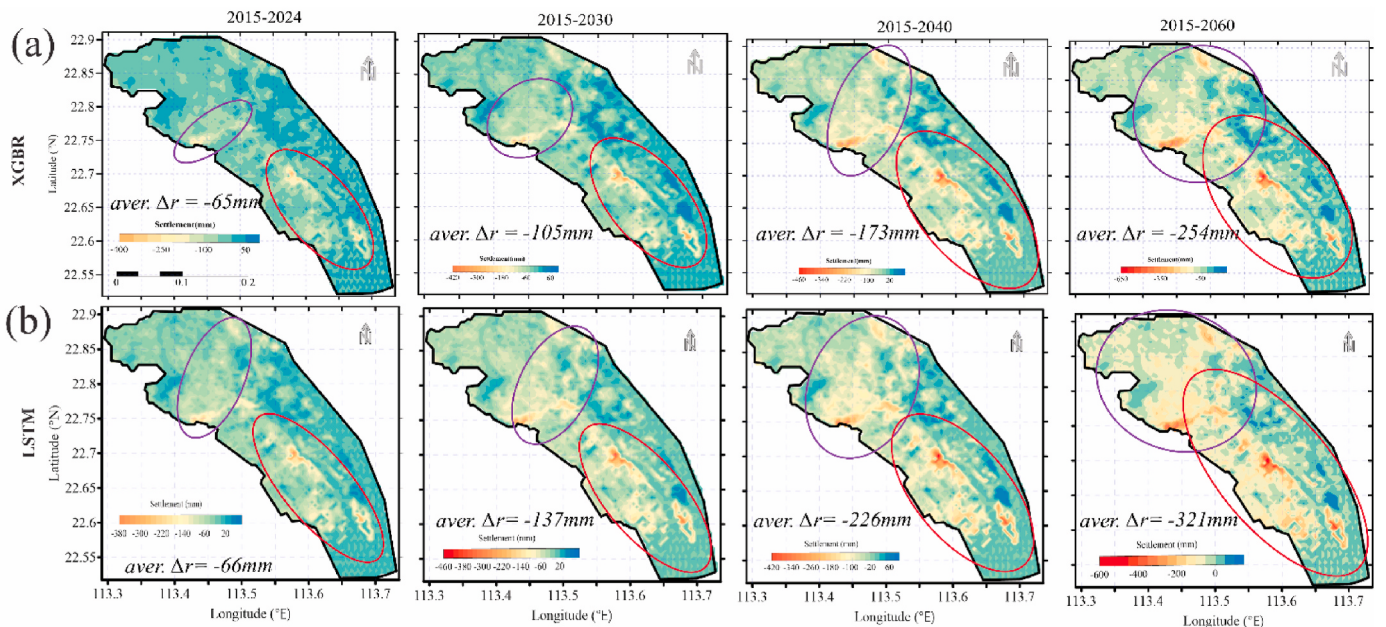
#### 4.2.3. Performance evaluation

Table 4 presents the performance metrics for two models, XGBR and LSTM, as applied to the LS simulation over the calibration, training, and validation/test. The coefficient of determination  $R^2$  of XGBR (0.89678) indicates a high level of explained variance in the training dataset. Its RMSE value, estimated at 0.378907, suggests that the average squared forecast error is moderate, indicating a reasonable level of prediction error. The MAE (0.34678) shows that the average magnitude of the errors in predictions is relatively low. The LSTM model shows a lower  $R^2$  (0.849872) than the XGBR model, indicating a slightly lower fit to the

**Table 4**

Results of XGBR and LSTM models for LS simulation.

| Period               | Performance measure | XGBR     | LSTM     |
|----------------------|---------------------|----------|----------|
| Calibration/Training | $R^2$               | 0.896780 | 0.849872 |
|                      | RMSE                | 0.378907 | 0.675438 |
|                      | MAE                 | 0.346780 | 0.567890 |
| Validation/Test      | $R^2$               | 0.902346 | 0.823678 |
|                      | RMSE                | 0.305678 | 0.556891 |
|                      | MAE                 | 0.456780 | 0.672340 |



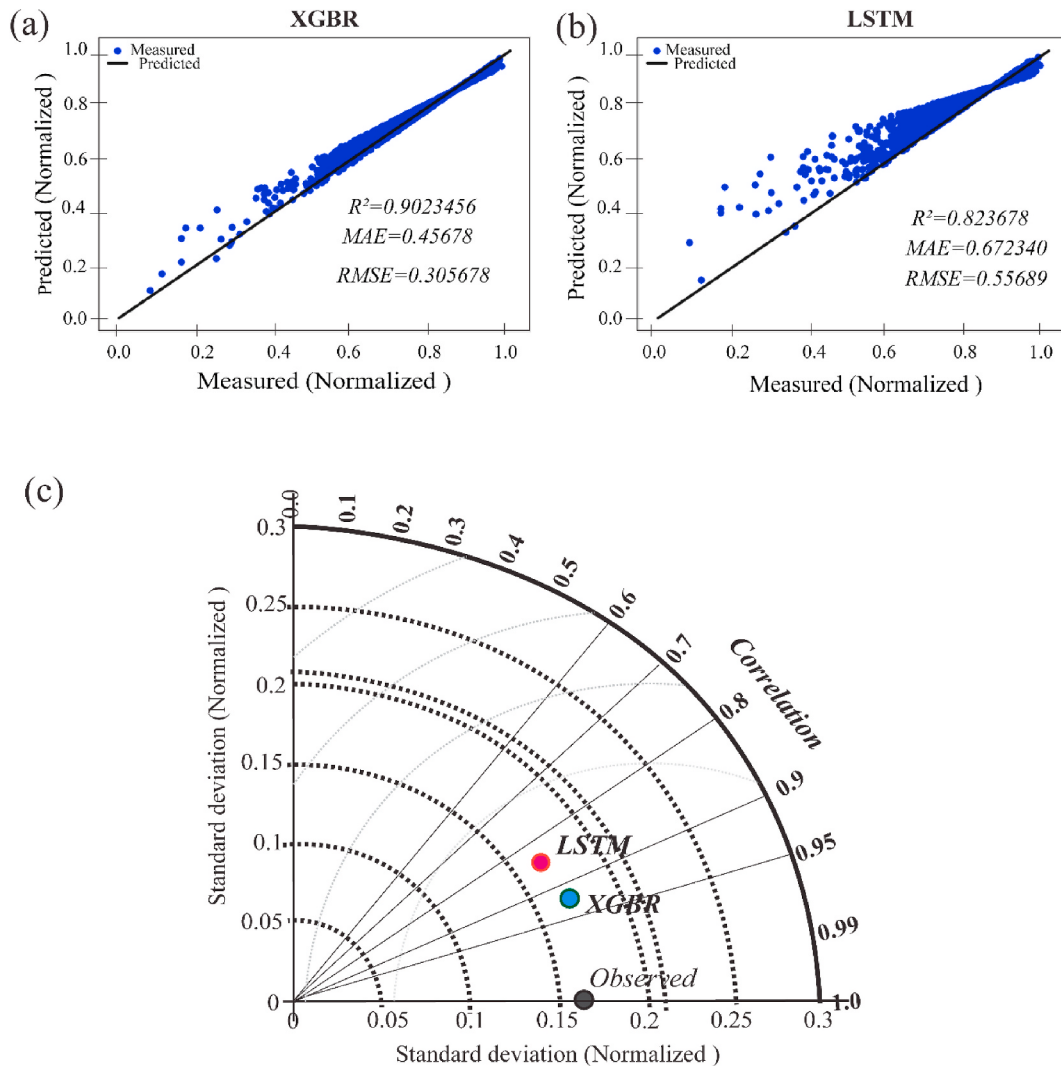
**Fig. 6.** LS simulation of 2024, 2030, 2040, and 2060 using  $\Delta r$ . a) Simulation with XGBR. b) Simulation with LSTM. LSTM predicts severe subsidence compared to XGBR to 2060.

training data. However, its RMSE is considerably higher at 0.675438, which suggests that the predictions are less accurate on average compared to XGBR. Likewise, the MAE is 0.567890, which is higher than XGBR's, indicating that the errors are larger on average.

During the testing phase, the RMSE, MAE, and  $R^2$  are used to measure the model's performance against the normalized GWL. Fig. 7a shows that XGBR has a lower RMSE value of 0.305678, which means that XGBR's predictions are pretty close to the real values. While the  $R^2 = 0.9023$  (close to 1) indicates that the model explains a majority of the variability in the response variable. Its MAE of 0.4568 shows a good average prediction error validated by a satisfactory value. Fig. 7b shows the LSTM performance with an RMSE of 0.55689. Its  $R^2 = 0.823678$  is still a strong score, but it's lower than XGBR, suggesting that LSTM doesn't capture as much of the variability. In addition, the average prediction error is larger than in XGBR, indicating less precise predictions on average with an MAE of 0.6723. Furthermore, the XGBR outperforms the LSTM model across all three performance metrics. The model shows improved performance from training to validation in terms of RMSE, which is a positive indicator of its generalization capabilities. The LSTM model's performance metrics are worse on the validation set than on the training set, which is a small sign that it might be overfitting the training data (Table 4).

#### 4.2.4. Taylor diagram analysis

The Taylor diagram (Fig. 7c) is a graphical summary of how closely a pattern (or set of patterns) matches observations. It is used to compare the skills of different models or to assess the performance of a model concerning observations based on the correlation coefficient and the standard deviation. The LSTM model's normalized standard deviation is almost the same as the observed data. This means that the model's predictions are quite variable, just like the observed data. It has a correlation with observed data that appears to be around 0.85, which suggests a very high degree of linear relationship between the model's predictions and the observed data. Besides, the normalized standard deviation for the XGBR model is slightly lower than that of the observed data, suggesting that the model's predictions are less variable than the observations. The correlation of the XGBR model with the observed data is slightly lower than that of the LSTM model, but still above 0.9, which is considered high and indicates a strong linear relationship. Both models exhibit high correlation coefficients with the observed data. XGBR is a little closer to the perfect correlation value of 1.0 than LSTM. This means that the predictions made by the XGBR model are more closely linked to the values that were observed. The data "observed" serves as a reference for the standard deviation, and both models are relatively close to this value, with XGBR almost identical and LSTM slightly lower. Both the correlation and standard deviation of XGBR and the observed LS pattern are a little closer to being in line with each other



**Fig. 7.** Model evaluation. XGBR (a) and LSTM (b) performance metrics evaluation limited with 3000 samples. The plot is performed using the groundwater level normalized values (GWL). evaluation. c) Taylor diagram.

than they are with the LSTM model. Overall, both models do a great job, but the XGBR model seems to be a little more in line with what we saw in terms of the pattern and variability of LS.

#### 4.2.5. LS risk prevention

Fig. 8 appears to offer a comparative analysis of LS forecasting between two scenarios. Scenario A corresponds to where the influence of GWL and BC is reduced by 80% and B otherwise. In scenario A (red frame), XGBR indicates that, with an 80% reduction in the significance of GWL and BC, there is a noticeable decrease in the areas and intensity of LS at around  $-111\text{ mm}$ . The purple outline presumably denotes areas most impacted by LS. Inside this outline, the coloration suggests less severe subsidence compared to scenario B "without reduction". Scenario A implies that mitigating the effects of groundwater extraction and regulating construction in areas prone to subsidence could significantly reduce the extent of future LS.

Moreover, in scenario A, the prediction of LSTM displays a more widespread and intense subsidence pattern, indicated by larger and more vivid red areas within the purple outline. The maps suggest that if current trends in GWL and BC continue unchecked, the LS by 2040 will be more severe ( $\Delta r = -226\text{ mm}$ ) than what is predicted under scenario A ( $\Delta r = -90\text{ mm}$ ). Comparing the real cumulated LS map from 2015 to 2022 (Fig. 4a) with the predicted scenarios, Fig. 8 indicates that the regions most affected by subsidence are likely to continue experiencing similar problems in the future if no intervention occurs. The 80% reduction in scenario A presents a future where proactive measures to lower GWL and regulate BC result in a markedly reduced severity of the LS issue. This suggests a path forward for land management and policy to mitigate subsidence risks. The "Risk Prevention" indicates that the

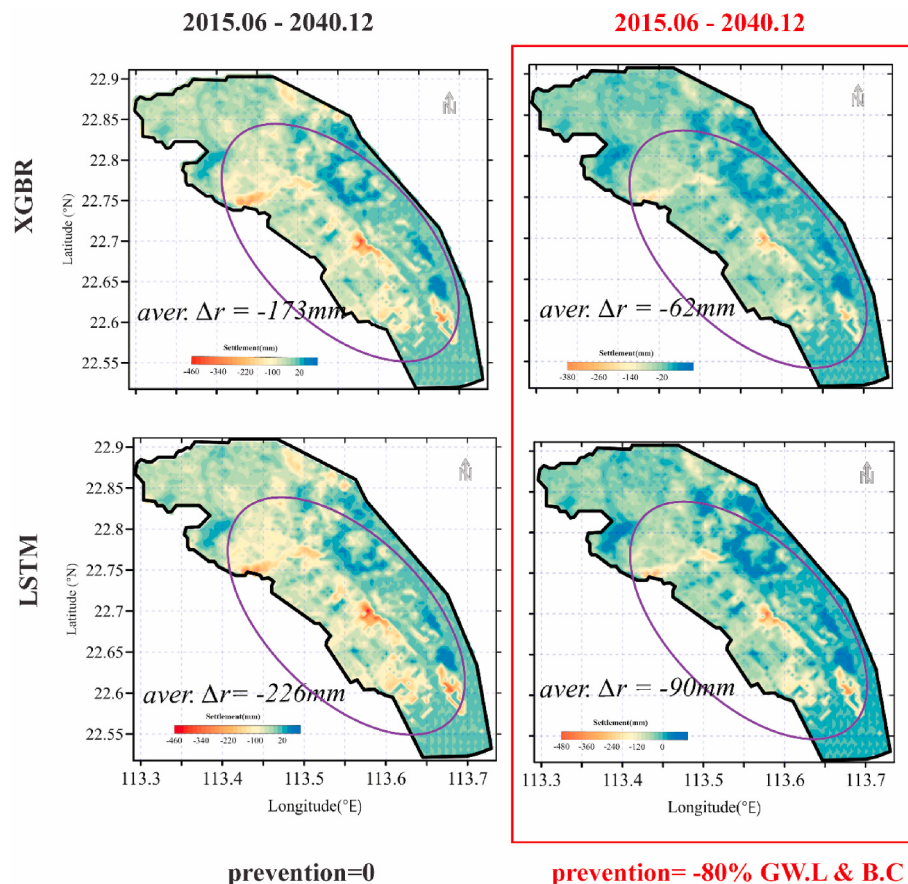
areas of greatest concern are effectively addressed when the impact of GWL and BC is reduced, underlining the potential benefits of such preventive measures. Thus, the LS risk prevention map provides a strong argument for proactive resource management and urban planning policies.

## 5. Discussion

In this discussion, we critically examine the impact and limitations of the delta rate ( $\Delta r$ ) approach in land subsidence (LS) prediction, while also highlighting its implications for the Nansha district. We first explore the effectiveness of the  $\Delta r$  approach, then address its constraints and the potential for future research to overcome these challenges. Finally, we discuss the practical significance of our findings for urban planning and policy-making, emphasizing the need for ongoing advancement in LS management.

### 5.1. Impact of using $\Delta r$ approach

In recent times, numerous researchers have devised various methodologies to address the challenges associated with LS sinking. For instance, in a study conducted by Rahmati et al. (2019b), a comparison was made between four tree-based machine learning models for LS hazard modeling in the Hamadan plain of Iran. The authors of this study employ the RF algorithm which exhibits a low predictive error. Based on their analysis, they determine that groundwater withdrawal emerges as the most relevant element contributing to LS. In the year 2020, Zamanirad et al. (2020) employed three machine learning models, namely boosted regression trees (BRTs), generalized additive model, and RF, in



**Fig. 8.** LS risk prevention using XGBR and LSTM. The red frame is scenario A: the result applied when reducing the GWL and BC to  $-80\%$   $\Delta r$  representativity in the data set. Scenario B is predicted LS in 2040 without applying any prevention. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

conjunction with four anthropological and geo-environmental predictors. The objective was to generate a spatial prediction map for an area in the southern region of Iran that is susceptible to LS. The study determined that the generalized additive model had the highest level of effectiveness as a susceptibility model within the designated study region. Based on the relative contribution test, it was determined that the primary predictive factor for LS occurrence is the decline of groundwater level, which accounts for 77.5% of the overall contribution. This quantitative analysis assesses the significance of several factors concerning LS. Moreover, Wang et al. (2023) propose a novel approach for simulating LS using a combination of Extremely Randomized Trees and the Monte Carlo algorithm from compressive layers. This approach enables the authors to do a quantitative analysis of the significance of various factors in the Beijing Plain. The researchers have provided evidence to support the notion that the groundwater level is the primary factor influencing land surface deformation, accounting for a significant proportion ranging from 67.6% to 81.8%. In addition to the aforementioned good outcomes, it is important to consider the impact of groundwater level, drawdown, or extraction as a significant contributing factor to LS in various regions. Our analysis further substantiated this claim. However, none of the aforementioned studies have attempted to utilize forecasting techniques for risk prevention. One contributing component to this situation is the limited availability of time scale data, as the collection of LS data and the acquisition of important variables data on an annual basis incur substantial expenditures.

The delta rate ( $\Delta r$ ) approach, as employed in this study, represents a significant advancement in LS prediction and risk management. This methodology allows for the collection of subsidence data at varying times, facilitating the forecasting of subsidence patterns years in advance and thereby enhancing risk prevention strategies. The strength of the  $\Delta r$  approach lies in its capacity to model LS with fewer data points while still capturing the essential dynamics of the phenomenon, a particularly crucial feature in contexts where continuous data collection is challenging or cost-prohibitive. By applying the  $\Delta r$  method with advanced machine learning models like XGBR and LSTM, which are adept at processing complex, non-linear relationships among various factors influencing LS, our study demonstrates that even limited data can yield reasonably accurate forecasts. This approach is invaluable in resource-limited settings and represents an innovative use of technology in environmental risk assessment.

Moreover, the  $\Delta r$  approach is supported by feature importance and Taylor analyses. The former identifies key drivers as understood by different models, while the latter serves as a visual tool to confirm the high statistical agreement of both models with observed data, thereby bolstering confidence in their predictions. By simulating scenarios where the impact of factors like groundwater level (GWL) and building concentration (BC) is reduced, our study showcases the potential effectiveness of targeted interventions. This aspect of our research provides practical insights for urban planners and policymakers, demonstrating how strategic actions could significantly alter the trajectory of LS in urban environments.

In essence, the integration of the  $\Delta r$  approach with XGBR and LSTM models offers a robust framework for LS prediction, enhancing our ability to foresee and manage risks associated with urban LS. This methodological innovation contributes to more sustainable and resilient urban development, underscoring the potential effectiveness of reducing groundwater extraction and ensuring suitable construction practices. It offers a beneficial tool for decision-makers addressing LS concerns, paving the way for informed policy decisions and mitigation strategies.

## 5.2. Limitation of $\Delta r$ and future research

While the  $\Delta r$  approach has demonstrated its effectiveness, it is crucial to acknowledge its inherent limitations, particularly the assumption that  $\Delta r$  remains constant over time in the absence of continuous data. This assumption may not always reflect the complex

and dynamic nature of real-world scenarios, where environmental and anthropogenic factors can cause fluctuations in LS rates. Such an oversimplification could lead to discrepancies between predicted and actual LS patterns. In our study, we attempted to mitigate this limitation by incorporating a range of variables and scenarios into the XGBR and LSTM models, thereby enhancing the robustness of our analysis and providing a more nuanced understanding of LS risks. This methodological approach, while it cannot entirely substitute for continuous, real-time data collection, offers a practical alternative in situations where the costs and logistics of extensive geoscience exploration are prohibitive. Furthermore, the predictive power of the models is strengthened by comparing their outputs with actual observed data, allowing for ongoing refinement and improvement of model accuracy. However, it's important to consider the nature of the dataset and the specific domain knowledge that may influence the predictive performance of certain features in one model over another. Variations in how different algorithms process feature interactions and non-linear relationships can also impact the ranking of feature importance. These considerations highlight the importance of a multi-faceted approach to model development and validation.

To address these challenges in future research, we recommend the development of more adaptive models that can account for temporal variations in  $\Delta r$  and integrate real-time data updates. Such advancements would enhance the predictive accuracy of LS forecasting models and provide a more reliable tool for urban planners and policymakers. Additionally, further exploration into the effects of various environmental and anthropogenic factors on LS, beyond the scope of our current dataset, would contribute to a deeper understanding of this complex phenomenon.

## 5.3. Implications for Nansha district

The findings of our study have profound implications for the Nansha district, a region grappling with the challenges of LS. Our analysis pinpoints GWL and BC as the primary drivers of LS in this area, as evidenced by the data presented in Fig. 5a. This crucial insight offers a strategic pathway for local authorities and urban planners: by effectively managing GWL and reducing BC, there's a potential to significantly curtail the subsidence rate. Such measures are not just theoretical projections but are grounded in robust statistical evidence, including a notable 80% targeted reduction in LS (−80% (GWL & BC)). This ambitious target, backed by our models' 70–95% confidence interval as depicted in Fig. 5b, provides a reliable foundation for policymakers to base their decisions upon. In practical terms, these findings equip urban planners and policymakers in Nansha with a powerful tool to forecast and mitigate LS risks. Implementing informed land management strategies based on our model's predictions could dramatically reduce LS-related damages, safeguarding the district's infrastructure and enhancing its overall sustainability and safety. This approach moves beyond conventional strategies, offering a data-driven path that not only anticipates future LS scenarios but also provides actionable insights for proactive urban development and risk management. The implications of this study, therefore, extend well beyond academic circles, offering tangible, actionable solutions for one of the most pressing urban challenges in the Nansha district.

## 6. Conclusions

This study marks a pivotal advancement in understanding and predicting land subsidence (LS), harnessing the power of delta-rate time-stamp calculus alongside cutting-edge machine learning models such as the eXtreme Gradient Boosting Regressor (XGBR) and Long Short-Term Memory (LSTM). Our comprehensive analysis reveals the significant influence of groundwater level (GWL) and building concentration (BC) on LS, with feature importance analysis underscoring their predominant roles. The robustness of our models is further validated through a Taylor

diagram, demonstrating a high correlation with observed data and affirming their predictive reliability. In a scenario where proactive measures are implemented to reduce GWL and BC, our study projects a notable decrease in LS by 2040. This forward-looking insight is crucial for environmentalists, urban planners, and policymakers, illustrating the transformative impact of informed policy changes and strategic interventions on mitigating LS risks. Such foresight is not just a theoretical exercise but a practical roadmap for sustainable urban development and the preservation of both natural and built environments.

As we reflect on the implications of our research, it becomes clear that the journey toward a deeper understanding and effective management of LS is ongoing. In this vein, we propose the following areas as critical pathways for future exploration:

- Real-time data integration: Future research should focus on incorporating real-time environmental data into LS models, enhancing their responsiveness and accuracy.
- Interdisciplinary collaboration: Encouraging collaboration across various fields like geology, urban planning, and environmental science can lead to more holistic approaches to LS challenges.
- Long-term socio-economic studies: Investigating the long-term impacts of LS on communities and economies will provide invaluable insights for policy formulation.
- Practical application in urban Planning: Bridging the gap between theoretical models and their practical application in urban development strategies is essential for real-world impact.

Our findings and these future directions offer a beacon of hope and guidance in the quest to tackle LS challenges. By continuing to innovate and expand our research horizons, we can pave the way for safer, more resilient urban landscapes, ensuring the well-being of communities and the integrity of our urban fabric for generations to come.

## Appendix

The Long Short-Term Memory (LSTM) model consists of a memory cell ( $C_t$ ) that is responsible for storing information, as well as three circular gates that control the flow of information within the LSTM cell (Fig.A1). The initial gate, as proposed by Gers et al. (2000) is referred to as the forget gate to regulate the amount to which the cell state vector  $C_{t-1}$  will be disregarded. The input, forget, and output gates in the internal LSTM model cell are represented by  $i$ ,  $f$ , and  $O$ , respectively. Furthermore, the variables  $h_t$  and  $C_t$  are used to denote the hidden state and the cell state at a given time  $t$ , respectively. The equations presented by Kratzert et al. (2018) provide a mathematical representation of the various gate and cell states of the LSTM model as

### • Input gate

$$i_t = \sigma(W_i X_i + U_i h_{t-1} + b_i) \quad (\text{A.1})$$

where  $i_t$ ,  $\sigma$ ,  $W_i$ ,  $X_i$ ,  $U_i$ ,  $h_{t-1}$ , and  $b_i$  represents the input gate vector (ranged between 0 and 1), sigmoidal function, weight connecting the input gate, weights from the input, output from the previous time step, and bias vector, respectively.

### • Forget gate

$$f_t = \sigma(W_f X_t + U_f h_{t-1} + b_f) \quad (\text{A.2})$$

where  $f_t$ ,  $W_f$ ,  $U_f$ , and  $b_f$  are the outputs of a vector forget gate with a value between 0 and 1 and a weight forget gate with inputs, input weights, and a bias vector, respectively.

### • Output gate

$$O_t = \sigma(W_o X_t + U_o h_{t-1} + b_o) \quad (\text{A.3})$$

where  $O_t$ ,  $W_o$ ,  $U_o$ , and  $b_o$  are the outputs of a vector output gate with a value between 0 and 1 and a weight forget gate with inputs, input weights, and a bias vector, respectively.

### • Cell state

## CRediT authorship contribution statement

**Jianxin Liu:** Writing – review & editing, Writing – original draft, Project administration, Investigation, Funding acquisition. **Wenxiang Liu:** Writing – original draft, Validation, Investigation, Data curation. **Fabrice Blanchard Allechy:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis. **Zhiwen Zheng:** Writing – original draft, Visualization, Resources, Formal analysis, Data curation. **Rong Liu:** Writing – review & editing, Supervision, Resources, Investigation, Funding acquisition. **Kouao Laurent Kouadio:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Conceptualization.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Data will be made available on request.

## Acknowledgments

The authors express their gratitude to the Guangdong Provincial Geological Bureau for the hydrogeology and engineering geology data as well as the provision of ground subsidence level measurement data. The research is supported by the National Natural Science Foundation of China (grant nos. 42130810, 41774149, and 42374093).

The possible update vector for the cell state is computed as follows using the last hidden state ( $h_{t-1}$ ) and current input ( $x_t$ ):

$$\bar{C}_t = \tanh(W_c X_t + U_c h_{t-1} + b_c) \quad (\text{A.4})$$

where  $C_t$  is the cell state at the previous time ranged between  $-1$  and  $1$ .  $\tanh$  indicates the hyperbolic tangent function. The cellular state ( $C_t$ ), as determined by the outcome of Equation A.4 is subsequently revised as

$$C_t = f_t^\circ C_{t-1} + i_t^\circ \bar{C}_t \quad (\text{A.5})$$

The computation of the new hidden state ( $h_t$ ) is achieved by integrating the outcomes of the output gate and the cell state.

$$h_t = \tanh(C_t) O_t \quad (\text{A.6})$$

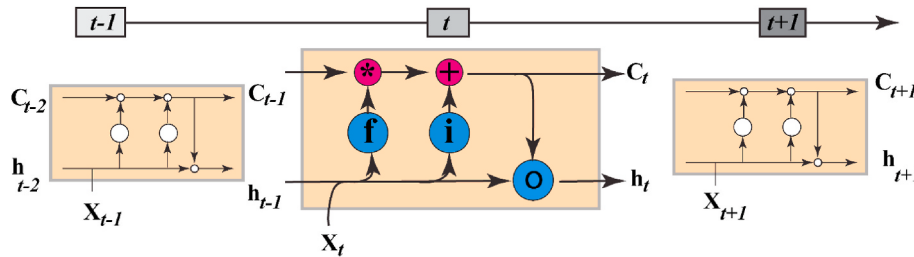


Fig. A.1. Architecture of Long-Short-Term Memory (LSTM)

## References

- Abidin, H.Z., Andreas, H., Gumilar, I., Fukuda, Y., Pohan, Y.E., Deguchi, T., 2011. Land subsidence of Jakarta (Indonesia) and its relation with urban development. *Nat. Hazards* 59, 1753–1771. <https://doi.org/10.1007/s11069-011-9866-9>.
- Bagheri-Gavkosh, M., Hosseini, S.M., Ataie-Ashtiani, B., Sohani, Y., Ebrahimian, H., Morovat, F., Ashrafi, S., 2021. Land subsidence: a global challenge. *Sci. Total Environ.* 778 <https://doi.org/10.1016/j.scitotenv.2021.146193>.
- Bai, L., Jiang, L., Wang, H., Sun, Q., 2016. Spatiotemporal characterization of land subsidence and uplift (2009–2010) over wuhan in central China revealed by terrasar-X insar analysis. *Rem. Sens.* 8, 350.
- Brown, S., Nicholls, R.J., 2015. Subsidence and human influences in mega deltas: the case of the Ganges–Brahmaputra–Meghna. *Sci. Total Environ.* 527, 362–374.
- Budhu, M., Adiyaman, I.B., 2010. Mechanics of land subsidence due to groundwater pumping. *Int. J. Numer. Anal. Methods Geomech.* 34, 1459–1478.
- Charpentier, A., James, M., Ali, H., 2022. Predicting drought and subsidence risks in France. *Nat. Hazards Earth Syst. Sci.* 22, 2401–2418.
- Chaussard, E., Wdowinski, S., Cabral-Cano, E., Amelung, F., 2014. Land subsidence in central Mexico detected by ALOS InSAR time-series. *Remote Sens. Environ.* 140, 94–106.
- Cigna, F., Tapete, D., 2022. Urban growth and land subsidence: multi-decadal investigation using human settlement data and satellite InSAR in Morelia, Mexico. *Sci. Total Environ.* 811, 152211 <https://doi.org/10.1016/j.scitotenv.2021.152211>.
- Corbeau, J., Gonzalez, O.L., Clouard, V., Rolandone, F., Leroy, S., Keir, D., Stuart, G., Momplaisir, R., Boisson, D., Prépetit, C., 2019. Is the local seismicity in western Hispaniola (Haiti) capable of imaging northern Caribbean subduction? *Geosphere* 15, 1738–1750.
- Deng, Z., Ke, Y., Gong, H., Li, X., Li, Z., 2017. Land subsidence prediction in Beijing based on PS-InSAR technique and improved Grey-Markov model. *GIScience & Remote Sens.* 54, 797–818.
- Du, Z., Ge, L., Ng, A.H.-M., Lian, X., Zhu, Q., Horgan, F.G., Zhang, Q., 2021. Analysis of the impact of the South-to-North water diversion project on water balance and land subsidence in Beijing, China between 2007 and 2020. *J. Hydrol.* 603, 126990.
- Ebrahimi, H., Feizizadeh, B., Salmani, S., Azadi, H., 2020. A comparative study of land subsidence susceptibility mapping of Tasuj plane, Iran, using boosted regression tree, random forest and classification and regression tree methods. *Environ. Earth Sci.* 79, 1–12.
- Ellis, J., Knight, J.E., White, J.T., Sneed, M., Hughes, J.D., Ramage, J.K., Braun, C.L., Teeple, A., Foster, L.K., Rendon, S.H., others, 2023. Hydrogeology, Land-Surface Subsidence, and Documentation of the Gulf Coast Land Subsidence and Groundwater-Flow (GULF) Model, Southeast Texas, 1897–2018. <https://doi.org/10.3133/pp1877>.
- Fernández-Torres, E.A., Cabral-Cano, E., Novelo-Casanova, D.A., Solano-Rojas, D., Havazli, E., Salazar-Tlacazani, L., 2022. Risk assessment of land subsidence and associated faulting in Mexico City using InSAR. *Nat. Hazards* 112, 37–55.
- Foroughnia, F., Nemati, S., Maghsoudi, Y., Perissin, D., 2019. An iterative PS-InSAR method for the analysis of large spatio-temporal baseline data stacks for land subsidence estimation. *Int. J. Appl. Earth Obs. Geoinf.* 74, 248–258.
- Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* 29, 1189–1232. <https://doi.org/10.2307/2699986>.
- Galloway, D.L., Burbey, T.J., 2011. Regional land subsidence accompanying groundwater extraction. *Hydrogeol. J.* 19, 1459.
- Gelete, G., 2023. Application of hybrid machine learning - based ensemble techniques for rainfall - runoff modeling. *Earth Sci. Informatics* 2475–2495. <https://doi.org/10.1007/s12145-023-01041-4>.
- Gers, F.A., Schmidhuber, J., Cummins, F., 2000. Learning to forget: Continual prediction with LSTM. *Neural Comput* 12, 2451–2471.
- Ghorbani, Z., Khosravi, A., Mojtahedi, F.F., Javadnia, E., Nazari, A., others, 2022. Use of InSAR Data for the Measurement of Land Subsidence in Response to Groundwater Fluctuations and Climate Change. Case Study: Ardabil Plain, Iran.
- Golian, M., Saffarzadeh, A., Katibeh, H., Mahdad, M., Saadat, H., Khazaei, M., Sametzadeh, E., Ahmadi, A., Sharifi Teshnizi, E., Samadi Darafshani, M., others, 2021. Consequences of groundwater overexploitation on land subsidence in Fars Province of Iran and its mitigation management programme. *Water Environ. J.* 35, 975–985.
- Guzy, A., Malinowska, A.A., 2020. State of the art and recent advancements in the modelling of land subsidence induced by groundwater withdrawal. *Water* 12, 2051.
- Hayati, N., Widodo, A., Kurniawan, A., Sanjiwani, I.D.M.A., Darminto, M.R., Yudha, I.S., Sumantyo, J.T.S., 2022. Small baselines techniques of time series InSAR to monitor and predict land subsidence causing flood vulnerability in Sidoarjo, Indonesia. *Geomatics, Nat. Hazards Risk* 13, 2124–2150.
- Ho, T.K., 1995. Random decision forests. In: *Proceedings of 3rd International Conference on Document Analysis and Recognition*, pp. 278–282.
- Hongdong, F., Kazhong, D., Chengyu, J., Chuanguang, Z., Jiqun, X., 2011. Land subsidence monitoring by D-InSAR technique. *Min. Sci. Technol.* 21, 869–872. <https://doi.org/10.1016/j.mstc.2011.05.030>.
- Koch, J., Schneider, R., 2022. Long short-term memory networks enhance rainfall-runoff modelling at the national scale of Denmark. *GEUS Bull.* 49.
- Kouadio, K.L., Kouame, L.N., Drissa, C., Mi, B., Kouamelan, K.S., Gnoleba, S.P.D., Zhang, H., Xia, J., 2022. Groundwater flow rate prediction from geo-Electrical features using support vector machines. *Water Resour. Res.* 58 <https://doi.org/10.1029/2021WR031623>.
- Kouadio, K.L., Liu, J., Kouamelan, S.K., Liu, R., 2023a. Ensemble learning Paradigms for flow rate prediction boosting. *Water Resour. Manag.* 37, 4413–4431. <https://doi.org/10.1007/s11269-023-03562-5>.
- Kouadio, K.L., Liu, J., Liu, R., 2023b. watex: machine learning research in water exploration. *SoftwareX* 22, 101367. <https://doi.org/10.1016/j.softx.2023.101367>.
- Kratzert, F., Klotz, D., Brenner, C., Schulz, K., Herrnegger, M., 2018. Rainfall-runoff modelling using long short-term memory (LSTM) networks. *Hydrol. Earth Syst. Sci.* 22, 6005–6022.
- Li, H., Zhu, L., Dai, Z., Gong, H., Guo, T., Guo, G., Wang, J., Teatini, P., 2021. Spatiotemporal modeling of land subsidence using a geographically weighted deep learning method based on PS-InSAR. *Sci. Total Environ.* 799, 149244.
- Li, Z., Chen, Q., Xue, Y., Qiu, D., Chen, H., Kong, F., Liu, Q., 2023. Numerical investigation of processes, features, and control of land subsidence caused by groundwater extraction and coal mining: a case study from eastern China. *Environ. Earth Sci.* 82, 82.
- Marfai, M.A., King, L., 2007. Monitoring land subsidence in Semarang, Indonesia. *Environ. Geol.* 53, 651–659. <https://doi.org/10.1007/s00254-007-0680-3>.
- Martin, A., Ashish, A., Paul, B., Eugene, B., Zhifeng, C., Craig, C., Greg, S.C., Andy, D., Jeffrey, D., Matthieu, D., Sanjay, G., Ian, G., Andrew, H., Geoffrey, I., Michael, I., Jia, Y., Rafal, J., Lukasz, K., Manjunath, K., Josh, L., Dandelion, M., Rajat, M.,

- Sherry, M., Derek, M., Chris, O., Mike, S., Jonathon, S., Benoit, S., Ilya, S., Kunal, T., Paul, T., Vincent, V., Vijay, V., Fernanda, V., Oriol, V., Pete, W., Martin, Wattenberg, Martin, Wicke, Yuan, Y., Xiaoqiang, Z., 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. <https://doi.org/10.5281/zenodo.10126399>.
- Mohammady, M., Pourghasemi, H.R., Amiri, M., 2019. Assessment of land subsidence susceptibility in Semnan plain (Iran): a comparison of support vector machine and weights of evidence data mining algorithms. *Nat. hazards* 99, 951–971.
- Motagh, M., Walter, T.R., Sharifi, M.A., Fielding, E., Schenk, A., Anderssohn, J., Zschau, J., 2008. Land subsidence in Iran caused by widespread water reservoir overexploitation. *Geophys. Res. Lett.* 35.
- Nishi, Katsunoshin, Kawai, M., Cahyono, B.E., Waqar, M.M., Nishi, Kaori, Tetuko Sri Sumantyo, J., 2023. Consecutive DInSAR and well based on the law of material conservation between land surface pressure and ground water to observe land subsidence. *Geocarto Int* 38, 2159069.
- Nourani, V., Elkiran, G., Abdullahi, J., 2019. Multi-station artificial intelligence based ensemble modeling of reference evapotranspiration using pan evaporation measurements. *J. Hydrol.* 577, 123958.
- Othman, A., Abotalib, A.Z., 2019. Land subsidence triggered by groundwater withdrawal under hyper-arid conditions: case study from Central Saudi Arabia. *Environ. Earth Sci.* 78, 1–8.
- Pacheco-Martinez, J., Hernandez-Marin, M., Burbey, T.J., González-Cervantes, N., Ortiz-Lozano, J.A., Zermeno-De-Leon, M.E., Solis-Pinto, A., 2013. Land subsidence and ground failure associated to groundwater exploitation in the Aguascalientes Valley, México. *Eng. Geol.* 164, 172–186.
- Pan, Y., Ding, H., Li, J., Shum, C.K., Mallick, R., Jiao, J., Li, M., Zhang, Y., 2022. Transient hydrology-induced elastic deformation and land subsidence in Australia constrained by contemporary geodetic measurements. *Earth Planet. Sci. Lett.* 588, 117556.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. PyTorch: an imperative Style, high-performance deep learning library. In: *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., pp. 8024–8035.
- Phien-wej, N., Giao, P.H., Nutalaya, P., 2006. Land subsidence in Bangkok, Thailand. *Eng. Geol.* 82, 187–201. <https://doi.org/10.1016/j.enggeo.2005.10.004>.
- Rahmati, O., Falah, F., Naghibi, S.A., Biggs, T., Soltani, M., Deo, R.C., Cerdà, A., Mohammadi, F., Tien Bui, D., 2019a. Land subsidence modelling using tree-based machine learning algorithms. *Sci. Total Environ.* 672, 239–252. <https://doi.org/10.1016/j.scitotenv.2019.03.496>.
- Rahmati, O., Golkarian, A., Biggs, T., Keesstra, S., Mohammadi, F., Daliakopoulos, I.N., 2019b. Land subsidence hazard modeling: machine learning to identify predictors and the role of human activities. *J. Environ. Manage.* 236, 466–480. <https://doi.org/10.1016/j.jenvman.2019.02.020>.
- Rajaei, T., Ebrahimi, H., Nourani, V., 2019. A review of the artificial intelligence methods in groundwater level modeling. *J. Hydrol.* 572, 336–351.
- Raju, A., Nanda, R., Singh, A., Malik, K., 2022. Multi-temporal analysis of groundwater depletion-induced land subsidence in Central Ganga Alluvial plain, Northern India. *Geocarto Int* 37, 11732–11755.
- Shahbazi, S., Mousavi, Z., Rezaei, A., 2022. Constraints on the hydrogeological properties and land subsidence through GNSS and InSAR measurements and well data in Salmas plain, northwest of Urmia Lake, Iran. *Hydrogeol. J.* 30, 533–555.
- Shi, L., Gong, H., Chen, B., Zhou, C., 2020a. Land subsidence prediction induced by multiple factors using machine learning method. *Remote Sens* 12, 4044.
- Shi, L., Gong, H., Chen, B., Zhou, C., 2020b. Land subsidence prediction induced by multiple factors using machine learning method. *Remote Sens* 12, 1–17. <https://doi.org/10.3390/rs12244044>.
- Shirzaei, M., Freymueller, J., Törnqvist, T.E., Galloway, D.L., Dura, T., Minderhoud, P.S. J., 2021. Measuring, modelling and projecting coastal land subsidence. *Nat. Rev. Earth Environ.* 2, 40–58. <https://doi.org/10.1038/s43017-020-00115-x>.
- Solorza, R., Carignano, C., Cioccale, M., Notarnicola, C., 2022. Ground Surface Subsidence in Córdoba, Argentina, revealed by multitemporal SAR interferometry. In: *2022 IEEE Biennial Congress of Argentina (ARGENCON)*, pp. 1–7.
- Strozzi, T., Wegmüller, U., Tosi, L., Bitelli, G., Spreckels, V., others, 2001. Land subsidence monitoring with differential SAR interferometry. *Photogramm. Eng. Remote Sensing* 67, 1261–1270.
- Tang, Y.Q., Cui, Z.D., Wang, J.X., Yan, L.P., Yan, X.X., 2008. Application of grey theory-based model to prediction of land subsidence due to engineering environment in Shanghai. *Environ. Geol.* 55, 583–593. <https://doi.org/10.1007/s00254-007-1009-y>.
- Tizzani, P., Berardino, P., Casu, F., Euillades, P., Manzo, M., Ricciardi, G.P., Zeni, G., Lanari, R., 2007. Surface deformation of long valley caldera and Mono basin, California, investigated with the SBAS-InSAR approach. *Remote Sens. Environ.* 108, 277–289.
- Tong, X., Sandwell, D.T., Smith-Konter, B., 2013. High-resolution interseismic velocity data along the San Andreas fault from GPS and InSAR. *J. Geophys. Res. Solid Earth* 118, 369–389.
- Van Niekerk, H.J., der Walt, I.J., 2006. Dewatering of the Far West Rand dolomitic area by gold mining activities and subsequent ground instability. *L. Degrad. & Dev.* 17, 441–452.
- Wang, Z., Guo, L., Gong, H., Li, X., Zhu, L., Sun, Y., Chen, B., Zhu, X., 2023. Land subsidence simulation based on Extremely Randomized Trees combined with Monte Carlo algorithm. *Comput. Geosci.* 178, 105415. <https://doi.org/10.1016/j.cageo.2023.105415>.
- Wolkersdorfer, C., Thiem, G., 1999. Groundwater withdrawal and land subsidence in northeastern Saxony (Germany). *Mine Water Environ* 18, 81–92.
- Xue, Y.-Q., Zhang, Y., Ye, S.-J., Wu, J.-C., Li, Q.-F., 2005. Land subsidence in China. *Environ. Geol.* 48, 713–720. <https://doi.org/10.1007/s00254-005-0010-6>.
- Yin, J., Yu, D., Yin, Z., Liu, M., He, Q., 2016. Evaluating the impact and risk of pluvial flash flood on intra-urban road network: a case study in the city center of Shanghai, China. *J. Hydrol.* 537, 138–145.
- Zamanirad, M., Sarraf, A., Sedghi, H., Saremi, A., Rezaee, P., 2020. Modeling the influence of groundwater exploitation on land subsidence susceptibility using machine learning algorithms. *Nat. Resour. Res.* 29, 1127–1141. <https://doi.org/10.1007/s11053-019-09490-9>.
- Zhou, C., Gong, H., Chen, B., Li, X., Li, J., Wang, X., Gao, M., Si, Y., Guo, L., Shi, M., others, 2019. Quantifying the contribution of multiple factors to land subsidence in the Beijing Plain, China with machine learning technology. *Geomorphology* 335, 48–61.
- Zhou, C., Lan, H., Gong, H., Zhang, Y., Warner, T.A., Clague, J.J., Wu, Y., 2020. Reduced rate of land subsidence since 2016 in Beijing, China: evidence from Tomo-PSInSAR using RadarSAT-2 and Sentinel-1 datasets. *Int. J. Remote Sens.* 41, 1259–1285.
- Zhu, L., Gong, H., Li, X., Wang, R., Chen, B., Dai, Z., Teatini, P., 2015. Land subsidence due to groundwater withdrawal in the northern Beijing plain. *China. Eng. Geol.* 193, 243–255. <https://doi.org/10.1016/j.enggeo.2015.04.020>.