UNIVERSIDADE FEDERAL DE MINAS GERAIS
School of Engineering
Graduate Program in Sanitation, Environment and Water Resources

Pedro Henrique Bernardes Solha

RESILIENT FLOOD MONITORING NETWORKS: Prioritizing Gauges with an eXplainable AI approach

Belo Horizonte
2026

Pedro Henrique Bernardes Solha

RESILIENT FLOOD MONITORING NETWORKS: Prioritizing Gauges with an eXplainable AI approach

Master's Thesis presented to the Graduate Program in Sanitation, Environment and Water Resources of the Universidade Federal de Minas Gerais, as a partial requirement for obtaining the title of Master in Sanitation, Environment and Water Resources.

Area of concentration: Water Resources

Line of research: Hydrological Processes Modeling

Supervisor: André Ferreira Rodrigues

Co-supervisor: Andrea Menapace

Belo Horizonte
2026

Solha, Pedro Henrique Bernardes.
S685r Resilient flood monitoring networks [recurso eletrônico] : prioritizing gauges with an eXplainable AI approach / Pedro Henrique Bernardes Solha. – 2026.
1 recurso online (115 f. : il., color.) : pdf.

Orientador: André Ferreira Rodrigues.
Coorientador: Andrea Menapace.

Dissertação (mestrado) – Universidade Federal de Minas Gerais, Escola de Engenharia.

Inclui bibliografia.

1. Engenharia sanitária – Teses. 2. Recursos hídricos – Desenvolvimento – Teses. 3. Inundações – Teses. 4. Monitoramento ambiental – Teses. 5. Inteligência artificial – Teses. 6. Piracicaba, Rio, Bacia (SP) – Teses. I. Rodrigues, André Ferreira. II. Menapace, Andrea. III. Universidade Federal de Minas Gerais. Escola de Engenharia. IV. Título.

CDU: 628(043)

Ficha catalográfica elaborada pelo bibliotecário Marcio A. A. Gomes CRB/6 2812
Biblioteca Prof. Mário Werneck, Escola de Engenharia da UFMG

UNIVERSIDADE FEDERAL DE MINAS GERAIS

Escola de Engenharia
Curso de Pós-Graduação em Saneamento, Meio Ambiente e Recursos Hídricos

FOLHA DE APROVAÇÃO

"Resilient Flood Monitoring Networks: Prioritizing Gauges With An Explainable Ai Approach"

PEDRO HENRIQUE BERNARDES SOLHA

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

Prof. Dr. ANDRÉ FERREIRA RODRIGUES - ORIENTADOR

Prof. Dr. Andrea Menapace - Coorientador

Prof. Dr. Antonio Manuel Herrera Fernandez

Prof. Dr. DANIELE DALLA TORRE

Aprovada pelo Colegiado do PG SMARH Versão Final pelos:

Coordenador Prof. Eduardo Coutinho de Paula
Orientador Prof. ANDRÉ FERREIRA RODRIGUES

Belo Horizonte, 16 de Janeiro de 2026.

Documento assinado eletronicamente por **Andre Ferreira Rodrigues**, **Professor do Magistério Superior**, em 16/01/2026, às 11:58, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Antonio Manuel Herrera Fernandez**, **Usuário Externo**, em 16/01/2026, às 12:30, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Daniele Dalla Torre**, **Usuário Externo**, em 16/01/2026, às 15:01, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Andrea Menapace**, **Usuário Externo**, em 17/01/2026, às 02:44, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Eduardo Coutinho de Paula**, **Coordenador(a) de curso de pós-graduação**, em 24/02/2026, às 10:21, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **4874269** e o código CRC **683937A7**.

Referência: Processo nº 23072.201557/2026-65 SEI nº 4874269

ACKNOWLEDGMENTS

I would first like to thank God for guiding me along the enlightened paths of my life.

I then thank the Brazilian people, especially the less favored portion of the population, who paid and still pay high taxes to provide public, free, and quality education to a small portion of the country's citizens.

I thank my mother, Maria Fernanda, for having gifted me with life and for having done everything she could so that I have access to the best education and receive the greatest affection in the world. I also thank my grandparents, Morethson and Maria do Carmo, for all the structure and support they gave to me and my mother throughout my life.

I thank Professor André Rodrigues, my supervisor, Professor Bruno Brentan, my unofficial co-supervisor, and Rodrigo Perdigão, my friend and colleague from PPG-SMARH, for the support, friendship, and companionship during the development of this research. I also thank the researcher Andrea Menapace, my co-supervisor, and the HYDRO-UAI research group, who were very important in the development of this work.

I thank the engineers and friends from the companies I worked for during my master's degree, HIDROBR and Knight Piésold, and also my colleagues from PPG-SMARH. I especially thank Ana Oliveira, Caio Voigt, Fabiano Ferreira, Felipe Laender, Guilherme Almeida, Lucas Furtado and Thaís Baêta, for the partnership and moments of joy during my double journey.

I also thank all my friends who accompanied my master's process and for the patience each of them had with my absences. Especially to Henrique Najar and to the Tomich brothers, Bernardo and Eduardo, for always being by my side.

I thank UFMG for the infrastructure and for everything that the graduate academic experience provided for my personal and professional growth.

Finally, I thank CAPES and CNPq for the financial resources granted during the research.

“This world is but a canvas to our imagination.”

Henry David Thoreau

RESUMO

O monitoramento e a previsão eficazes de cheias são fundamentais para os modernos Sistemas de Alerta Antecipado de Inundações (FEWS), especialmente à medida que as mudanças climáticas e a urbanização intensificam a vulnerabilidade hidrológica, aumentando a necessidade de estruturas de modelagem de cheias mais adaptativas. Embora modelos de Aprendizado de Máquina (ML), como o Perceptron Multicamadas (MLP) utilizado nesta pesquisa, ofereçam ferramentas poderosas para previsões rápidas e precisas do nível d'água, sua efetividade operacional é frequentemente comprometida por dois principais desafios: o aspecto operacional das redes de monitoramento, sensíveis à qualidade dos dados, e a natureza "caixa-preta" dos próprios modelos ML. Diante disso, esta pesquisa investiga uma metodologia integrada que utiliza Inteligência Artificial Explicável (XAI) não apenas para aumentar a transparência dos modelos, mas também como ferramenta para avaliar a resiliência da rede e priorizar estações de monitoramento críticas. A metodologia é aplicada à Bacia do Rio Piracicaba, em São Paulo, Brasil, com foco na cidade de Piracicaba, onde as inundações são um problema histórico em meio ao crescimento urbano. Para prever o nível d'água em horizontes de 3 a 24 horas, foi criado um conjunto de modelos MLP, empregando dados de 16 estações hidrometeorológicas. Um algoritmo de deterioração de dados que simula a falha progressiva e contínua de sensores foi utilizado para testar a resiliência da rede (perda de dados de 10% a 100%), quantificando o impacto por meio do coeficiente de Eficiência de Nash-Sutcliffe (NSE). Simultaneamente, a metodologia SHAP (SHapley Additive exPlanations) foi aplicada para elucidar algumas das decisões do modelo, medindo a contribuição de cada estação nas previsões e oferecendo uma fundamentação para a interpretação hidrológica. O alinhamento entre os testes empíricos de deterioração e as interpretações baseadas em XAI fornece evidências de que o modelo conseguiu aprender processos hidrologicamente coerentes. Para previsões de curto prazo (3 horas), a precisão do modelo mostrou-se mais sensível à falha da estação imediatamente a montante (713), refletindo a dinâmica de propagação da cheia rio abaixo. Para horizontes mais longos (24 horas), a importância deslocou-se para estações de meia bacia (55), indicando que o modelo capturou a física do tempo de propagação da onda de cheia. Outro achado emergiu ao testar uma rede reduzida composta apenas pelas cinco estações mais importantes: embora precisa com menos estações, essa rede "otimizada" mostrou-se criticamente frágil, refletindo o fato de que estações consideradas redundantes se revelam fundamentais para a resiliência de todo o sistema. O trabalho introduz uma nova abordagem que vai além da tradicional modelagem de acurácia, focando em uma avaliação mais profunda e voltada à resiliência de redes de monitoramento hidrometeorológico. Ao combinar testes empíricos de estresse com métodos de XAI, a intenção é oferecer uma abordagem mais clara e fundamentada em dados para a manutenção estratégica e a priorização dos recursos operacionais, aumentando a confiabilidade dos EWS em um contexto hidroclimático cada vez mais incerto.

Palavras-chave: Previsão de cheias, IA Explicável, Resiliência, Sistemas de Alerta, Redes de Monitoramento.

ABSTRACT

Effective flood monitoring and forecasting are fundamental to modern Flood Early Warning Systems (FEWS), particularly as climate change and urbanization intensify hydrological vulnerability, driving the need for more responsive flood-modeling frameworks. While Machine Learning (ML) models, such as the Multi-Layer Perceptron (MLP) used in this research, offer powerful tools for rapid and accurate stage-level prediction, their operational effectiveness is often undermined by two main challenges: the operational side of monitoring networks, which are sensitive to data quality, and the inherent "black-box" nature of the models themselves. Therefore, this research further investigates an integrated framework that leverages eXplainable AI (XAI) not only to enhance model transparency but also as a tool to assess monitoring network resilience and prioritize critical monitoring gauges. The methodology is applied to the Piracicaba River Basin, São Paulo, Brazil, focusing on the flood-prone city of Piracicaba. A set of MLP models was developed to forecast water level at lead-times ranging from 3 to 24 hours, using data from 16 hydro-meteorological stations. Monitoring network resilience was tested by applying a data deterioration algorithm on the MLP prediction models, that simulates progressive, continuous sensor failures (10% to 100% data loss), with the impact quantified by the Nash-Sutcliffe Efficiency (NSE). Concurrently, SHAP (SHapley Additive exPlanations) was employed to unveil part of the model's "black box," quantifying each gauge's contribution to the forecasts and providing a basis for hydrological interpretation. The complementarity between the empirical deterioration tests and the XAI-based interpretations provides evidence that the model could learn hydrologically coherent processes. For shorter lead times (3 hours), the model's forecast accuracy was most sensitive to the failure of the nearest upstream gauge, reflecting downstream flood routing dynamics. For longer lead times (24 hours), importance shifted to mid-basin gauges, indicating the model captured the physics of flood wave travel time. Another finding emerged from testing a reduced network of only the five most important gauges: while accurate using less gauges, this "optimized" network was critically fragile, reflecting the fact that gauges perceived as redundant prove critical to system-wide resilience. This research provides a novel and practical framework for moving beyond simple model accuracy modeling to a deeper, resilience-oriented assessment of flood monitoring networks. By integrating systematic empirical stress-testing with XAI techniques, it aims to deliver a more transparent, data-driven methodology to guide the strategic maintenance and prioritization of monitoring assets, ultimately enhancing the reliability of EWS in an increasingly uncertain hydro-climatic setting.

Keywords: Flood Forecasting, Explainable AI, Resilience, Warning Systems, Gauging Networks.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AI - Artificial Inteligence

ANN - Artificial Neural Network

CNN - Convolutional Neural Network

DL - Deep Learning

FEWS - Flood Early Warning System

GNN - Graph Neural Network

KGE - Kling-Gupta Efficiency

LSTM - Long Short-Term Memory

ML - Machine Learning

MLP - Multi-Layer Perceptron

MSE - Mean Squared Error

NSE - Nash-Sutcliffe Efficiency

NWP - Numerical Weather Prediction

PINN - Physics-Informed Neural Network

ReLU - Rectified Linear Unit

UNISDR - United Nations Office for Disaster Risk Reduction

XAI - Explainable AI

SUMÁRIO

1 INTRODUCTION

Since the turn of the millennium, hydrological risk has evolved rapidly due to the dual pressures of accelerating urbanization and shifting climate patterns. Global climate change is intensifying the hydrological cycle, leading to an increased frequency and magnitude of extreme events in many regions (Kundzewicz et al.,2013; IPCC, 2021). Simultaneously, the expansion of urban centers increases impervious surfaces, which reduces infiltration and favors runoff, amplifying flood peaks even under normal rainfall conditions (Miller & Hutchins, 2017). This confluence of factors has challenged historical assumptions that hydrological systems retain their statistical properties constant over time. In this era of non-stationarity beliefs, there is a need for adaptive, forward-looking approaches to flood risk management that can tackle properly the more uncertain hydrological future.

In response to this challenge, some solutions moved from a reactive focus on "flood protection" towards a more proactive framework of "flood risk management" (Plate, 2002) by the observed increase in human and economic losses (Dottori et al, 2018). This modern approach prioritizes mitigating the consequences of flooding, with the Flood Early Warning Systems (FEWS) serving as a central non-structural measure. An effective FEWS is designed to provide timely and actionable information to save lives and reduce economic damages to flood prone communities (UNISDR, 2009). The technical core of any modern FEWS is its monitoring and forecasting component, which relies on a network of hydro-meteorological sensors to provide the real-time data necessary to predict flood events and accurate hydrological modelling.

The forecasting modeling of hydrological systems have two main branches: process-based and data-driven approaches. While traditional process-based models offer physical interpretability, their field data necessity and assumptions often limit their utility in real-time operational contexts (Cea & Costabile, 2022). In the close past, data-driven methods, particularly those relying on Machine Learning (ML), have emerged as a powerful alternative, with some Deep Learning (DL) architectures like Long Short-Term Memory (LSTM) networks consistently demonstrating high predictive accuracy (Slater et al., 2025). However, the widespread adoption of these high-performing models has been criticized by their inherent “black box” nature. This lack of transparency hinders trust, making it harder to identify the cause of model failures or

confidently act on predictions (Thakker et al., 2020). This challenge has urged the adoption of eXplainable Artificial Inteligence (XAI) approaches, a field dedicated to making data-driven models more understandable (Gunning et al., 2019). In the context of resilient monitoring networks, XAI is a tool to tackle the research gap by moving beyond simple performance metrics. It serves as a diagnostic method to identify how the model's reliance on specific gauges is driven by hydrological processes rather than just statistical correlations.

However, even a transparent model remains operationally vulnerable if it relies on a fragile flow of information or data source. Although the data-driven methods in engineering are rising in popularity (Montáns et al., 2019), it has also introduced a dependency on data quality. Data-driven algorithms are bound by the information quality of their training sets (Li et al., 2023), lacking the physical intuition to identify corrupted inputs. In the absence of explicit redundancy to cross-verify readings, a 'black box' model may treat sensor errors as valid hydrological signals. This results in plausible-looking but incorrect predictions, hindering the ability of operators to perform the fault diagnosis necessary for a resilient monitoring network (Jung & Axelsson, 2024).

This leads to the central problem addressed by this research, which lies at the intersection of two critical challenges: the need for transparency in data-driven models and the operational reality that the monitoring networks data are vulnerable and prone to failure (Nguyen et al., 2021). This vulnerability is critical, since the reliability of the entire warning chain based on data-driven models is fundamentally dependent on the integrity and performance of the gauging network, without model reliance on coherent physical processes but rather on statistical correlations (Kirchner, 2006; Maier et al., 2024). While XAI has been used to interpret those models behaviors under ideal data conditions, there is a gap for a framework that uses it as a diagnostic tool to systematically assess the resilience of a forecasting system to the inevitable reality of network failures. The hypothesis driving this work is that a network prioritized based on hydrological consistency, validated through an XAI-integrated framework, will exhibit superior resilience compared to one based solely on statistical performance. Thus, this research moves beyond the question of "how accurate is the model?" to address a more operational question: "how robust is the warning system when its

components fail, and which components could be the most critical to its resilience?" This work proposes to contribute to the state of the art by extending the methodology proposed by Menapace et al. (2025), integrating progressive data deterioration with XAI to assess network resilience and provide a evidence-based framework for prioritizing critical monitoring gauges.

2 OBJECTIVES

2.1 General Objective

To evaluate the resilience of a data-driven FEWS to monitoring network failures by means of an explainable AI approach.

2.2 Specific Objectives

- To quantify the impact of missing input data on model accuracy using a time-series deterioration algorithm and performance metrics.
- To identify critical monitoring gauges by correlating XAI-based feature importance with performance degradation under sensor failure, highlighting gauges with the greatest impact on prediction accuracy.
- To assess the hydrological significance of model predictions to improve transparency and support decision-making in flood forecasting.

3 LITERATURE REVIEW

3.1 Early Warning Systems

The traditional approach to managing flood events has historically been centered on the paradigm of flood protection, a reactive strategy focused on containing floodwaters through structural defenses like dikes and levees. However, a series of catastrophic floods over the last several decades has demonstrated that absolute protection is both unachievable and unsustainable (Schanze, 2006). This recognition has catalyzed a significant paradigm shift towards a more proactive and holistic framework of flood risk management (Plate, 2002). This modern approach accepts that some level of residual risk is inevitable and therefore broadens the focus from solely preventing inundation to mitigating the adverse consequences for human health, the environment, and economic activity.

Within this comprehensive risk management framework, non-structural measures have gained prominence, and among the most critical of these is the FEWS. This type of warning system is not merely a technical tool but an integrated network of people and processes designed to prevent loss of lives and reduce property damage by providing timely and effective information (Collins & Kapucu, 2008). The UNISDR defines a complete and effective, people-centered FEWS as being built upon four essential and interlinked components: (1) Risk Knowledge; (2) Monitoring and Forecasting; (3) Warning Dissemination and Communication; and (4) Response Capability (UNISDR, 2009). The socio-economic benefits of such systems are significant, with studies demonstrating that effective warnings can substantially reduce flood damages, although this is highly dependent on factors like warning lead time and the public's capacity to take protective action (Parker et al., 2007).

The Monitoring and Forecasting component, which forms the technical core of the system, has undergone a significant technological evolution. Historically, monitoring relied on manual readings from staff gauges, which provided valuable but often delayed data. The advent of weather radar represented a major leap forward, providing the ability to estimate rainfall intensity over large areas with high spatial and temporal resolution (Mimikou & Baltas, 1994). While radar presents its own challenges, such as calibration and signal interference, it has become a cornerstone of modern forecasting, especially for flash floods where lead times are short (Binetti et al., 2022). This evolution has continued with the development of automated, real-time sensor networks. Modern systems, such as the one described by Sunkpho & Ootamakorn (2011), utilize a network of wireless sensors to monitor water level, streamflow, and precipitation, transmitting data in real-time via telemetry to a central server. This continuous stream of high-resolution data from multiple sources is the essential input that powers the sophisticated forecasting models at the heart of any modern FEWS. Recent systems increasingly leverage machine and deep learning algorithms to enhance prediction capabilities and integrate with advanced flood alert systems (Hayder et al. (2023).

3.2 Methodological Advances in Hydrological Forecasting

The forecasting engine at the core of a modern FEWS has evolved significantly over the past decades, largely following two parallel but distinct methodological paths:

process-based modeling and data-driven modeling (Gauch et al., 2020). The choice between these approaches often involves a trade-off between physical interpretability and predictive accuracy.

3.2.1 Process-based models

Process-based models represent the traditional approach to hydrological forecasting. These models attempt to explicitly simulate the physical processes of the land phase of the hydrological cycle, such as infiltration, runoff, and channel routing, based on fundamental physical laws like the conservation of mass and energy. They range in complexity from simpler, lumped conceptual models, which use simplified reservoir representations of basin storage, to complex, physically-based distributed models that solve differential equations (e.g., the Saint-Venant equations for shallow water flow) over a discretized grid representing the basin (Cea & Costabile, 2022).

The most prominent strength of process-based models relies on their physical interpretability and stakeholder visibility. Because their parameters are intended to represent tangible physical properties (e.g., soil type, land use and occupation, water infiltration, groundwater storage), they offer a way to explain which mechanisms are driving the basin's response. This makes them valuable tools for exploring what-if scenarios, such as the potential impacts of land-use change or different climate futures (Yuan et al., 2015).

However, in an operational forecasting context, process-based models face some limitations. They are notoriously data-intensive, often requiring extensive, high-quality field data on meteorology, topography, soil properties, and land cover for proper calibration and validation (Cea et al., 2010). Furthermore, they could be computational expensive especially for large-scale, high-resolution applications or for systems that require rapid, ensemble-based forecasts, limiting the application of such modeling in the FEWS context. The forecasting chain often begins with meteorological inputs from Numerical Weather Predicion (NWP) models or rainfall nowcasting systems, which have their own inherent uncertainties that propagate through the hydrological model, adding another layer of complexity to the modeling (De Luca et al., 2025; Yuan et al., 2015).

3.2.2 Machine Learning models

In parallel with advances in process-based physical models, recent years have witnessed a surge in the application of data-driven methods for engineering (Montáns et al., 2019). In particular, ML has become increasingly prominent in flood forecasting (Kuhaneswaran et al., 2025), enabled by the exponential rise in computational power and the widespread availability of monitoring data (Montáns et al., 2019). However, this abundance often introduces significant data redundancy, where vast amounts of collected information contribute little to model performance (Li et al., 2023). While some redundancy is necessary for system robustness (Jung and Axelsson, 2024), recent studies suggest that exploiting this redundancy is essential for optimizing computational efficiency (Chen et al., 2023).

This approach can have some advantages in an FEWS operational context. ML models are generally less demanding in terms of the variety of input data and specific domain knowledge they require, and their computational cost during inference (i.e., making a prediction) could be lower than that of complex physical models, making them well-suited for real-time and short-term applications (Muñoz et al., 2021). The Artificial Neural Networks (ANNs), and specifically the MLP adopted for the modeling in this research, has been a foundational tool. However, the field has advanced to embrace more sophisticated architectures capable of handling the temporal dependencies inherent in hydrological time-series (Tripathy & Mishra, 2024) and also within broader modeling frameworks to leverage diverse data sources and model structures (Frame et al., 2025). Models like the LSTM network have become the standard for many sequential data problems. Indeed, deep learning approaches have demonstrated capability for predicting hydrological extremes, sometimes outperforming traditional physical and statistical methods (Frame et al, 2022).

Recent research continues to push the boundaries of what is possible in the field of data-driven models for hydrology. State-of-the-art models now incorporate even more advanced architectures, such as Convolutional Neural Networks (CNNs) for capturing spatial patterns in rainfall, Attention Mechanisms, and Transformers, which have shown exceptional performance in long-sequence forecasting (Xu et al., 2023; Zhao et al., 2025). Furthermore, techniques like Graph Neural Networks (GNNs) are being

explored to explicitly incorporate the spatial connectivity of the river network into the model architecture (Liu et al., 2023).

Despite their high predictive accuracy, the primary challenge hampering a widespread adoption of ML models in critical, high-stakes decision-making is their "black box" nature. Because these models learn representations of the system that are not directly tied to physical principles or variables, it can be difficult for the government specialists, stakeholders and water managers to understand how the forecasting is being made. This lack of transparency hinders trust and makes it difficult to explain a model missed forecast for extreme events that may be outside the range of the training data. This has led to a growing field of research that focus on attempts to produce physically plausible and consistent ML models results (Kingston et al., 2005) and the development of Physics-Informed Neural Networks (PINNs) frameworks that aim to blend the predictive power of data-driven models with the physical constraints of process-based models (Bhasme et al., 2022).

3.3 Network Integrity and Data Quality

While the methodological sophistication of forecasting models has advanced significantly, their performance is fundamentally constrained by the quality and availability of the input data they rely upon. The most advanced model, whether process-based or data-driven, will produce unreliable forecasts if it is fed with incomplete or inaccurate data. This section reviews the literature on the operational challenges of maintaining the integrity of monitoring networks and the documented impact of data deterioration on model performance.

3.3.1 Sensor Failure and Maintenance in Monitoring Systems

Hydro-meteorological monitoring networks are one of the critical parts of any FEWS chain, yet they are inherently vulnerable to a range of failures. In real-world operational settings, sensors are exposed to harsh environmental conditions, leading to data gaps from technical malfunctions, power supply interruptions, or physical damage during the very flood events they are designed to monitor (Nguyen et al., 2021). Furthermore, data transmission via telemetry systems can be disrupted, leading to communication losses that create gaps in the real-time data stream. Quality assurance and control are

therefore critical, continuous processes required to identify and correct for anomalies or errors in the data before they are used for forecasting (Branisavljević et al., 2009).

The design, installation, and ongoing maintenance of these networks represent a significant logistical and financial challenge for the agencies responsible. The high costs associated with both capital investment and regular maintenance mean that networks are often sparse, particularly in developing regions or remote areas. This reality has led to a field of research focused on the optimal design of monitoring networks. Given that resources are finite, a critical question arises: what is the most cost-efficient number, type, and positioning of sensors required to achieve a target level of forecasting accuracy (Grimaldi et al., 2024). This line of research moves beyond simply adding more sensors and focuses on maximizing the informational value of each gauge. Recent work in snow hydrology, for instance, has shown that strategically monitoring localized hotspots, areas with untapped predictive power, can improve water supply forecasts more effectively than expensive, basin-wide mapping. This research demonstrates that the greatest gains in prediction come from measuring in the right places, not necessarily everywhere (Raleigh et al., 2025). This concept of informationally-rich hotspots is directly analogous to the challenge of identifying the most critical gauges in a flood monitoring network, a concept central to designing parsimonious yet effective FEWS. Recent studies, such as Menapace et al. (2025), have demonstrated how interpretable machine learning can be employed to formalize this prioritization, providing a data-driven framework for identifying the most influential sensors within a network.

3.3.2 The Impact of Data Deterioration on Model Performance

The inevitability of sensor failure means that missing values are a common problem in hydrological time-series. The process of filling these gaps, known as data imputation, has been extensively studied. A variety of techniques exist, ranging from simple statistical methods like interpolation to more complex data-driven approaches using machine learning (Kornelsen & Coulibaly, 2014). While these methods can provide reasonable estimates, they all introduce a degree of uncertainty and can never perfectly replicate the true, unobserved data (Gauch et al., 2025). Managing uncertainties is a persistent challenge in flood forecasting, particularly when using

data-driven approaches, which can be sensitive to input variations and model structure (Han et al., 2007).

The impact of this data scarcity on model performance is a significant concern. The challenge of prediction in ungauged basins has been a major research focus in hydrology for decades, aiming to produce reliable forecasts in locations with little to no historical data (Yanto et al., 2017). This problem is conceptually similar to forecasting with a deteriorated network, as both scenarios force models to operate with incomplete information. Recent studies have shown that ML models, particularly when combined with global precipitation datasets, can provide reliable flood predictions even in ungauged areas, demonstrating their potential to overcome some of the limitations of data scarcity (Nearing et al., 2023). Recent studies have shown that ML models combined with global precipitation datasets, can provide reliable flood predictions even in ungauged areas, demonstrating their potential to overcome some of the limitations of data scarcity (Rasheed et al., 2024). More advanced techniques like data assimilation offer a formal framework for integrating real-time observations with model predictions to correct for model errors and update the system's state, providing a robust method for handling imperfect and incomplete data streams (Zhang et al., 2024). Collectively, this body of research underscores a critical point: the performance and reliability of any flood forecasting model are directly tied to the integrity of its input data, and the ability to function effectively under conditions of data deterioration is a key measure of a system's operational resilience.

3.4 XAI for Understanding BlackBox Models

The black box drawback inherent in many high-performing ML models poses a significant barrier to their adoption in high-stakes fields like flood risk management, where decisions have direct consequences for public safety and infrastructure (Thakker et al., 2020). When a model cannot explain *why* it produced a certain forecast, it becomes difficult for human experts to trust its predictions, diagnose failures, or responsibly manage the system. This aligns with the long-standing call within hydrology to ensure models are 'getting the right answers for the right reasons, as stated by Kirchner (2006). This challenge has given rise to the rapidly growing field of XAI, which aims to develop methods that render the decisions and predictions of complex models understandable to humans (Gunning et al., 2019).

3.4.1 Principles of XAI

The field of XAI distinguishes between two fundamental concepts: interpretability and explainability (Adadi & Berrada, 2018). Interpretability refers to the extent to which a model's internal mechanics can be understood by a human. Simpler models like linear regression are inherently interpretable because the relationship between inputs and outputs is transparent. In contrast, complex models like deep neural networks are generally less interpretable. Explainability, on the other hand, refers to the ability to generate a human-understandable description of a model's decision-making process, even if the model itself is a black box. This is the primary goal of XAI (Samek et al., 2019).

XAI techniques are often categorized as either model-specific or model-agnostic. Model-specific methods are designed for a particular class of models (e.g., decision trees), while model-agnostic methods can be applied to any machine learning model, regardless of its internal structure (Ribeiro, Singh & Guestrin, 2016). This flexibility has made model-agnostic techniques particularly popular. One of the most prominent of these is SHAP, which is used in this research. The SHAP methodology is based on the concept of Shapley values from cooperative game theory, which provides a principled way to fairly attribute the contribution of each input feature to the outcome of the model's prediction. For each prediction, SHAP assigns a value to each input feature that represents its marginal contribution to pushing the prediction away from a baseline value, providing a powerful, theoretically grounded method for understanding feature importance (Lundberg & Lee, 2017).

3.4.2 Applications of XAI in Hydrology

The application of XAI in hydrology is a relatively new but rapidly expanding area of research, driven by the need to ensure that powerful ML models are not only accurate but also hydrologically plausible and trustworthy (Slater et al., 2025). Researchers have begun to apply XAI techniques to "peek inside the black box" of hydrological ML models to achieve several key objectives.

First, XAI has been used to validate how models are learning physically meaningful relationships. For instance, studies have used XAI to confirm that models correctly

identify antecedent precipitation and soil moisture as key drivers of runoff, aligning the model's learned behavior with established hydrological principles (Başağaoğlu et al., 2022). This helps to ensure that a model is getting the right answer for the right reasons, a crucial step in building confidence that it will generalize to unseen conditions (Kingston, Maier & Lambert, 2005).

XAI revealing which input features a model is relying on most heavily, hydrologists can identify when a model might be learning spurious correlations from the data or exhibiting unexpected behavior. Furthermore, XAI is being applied to solve specific operational challenges, such as found in Menapace et al. (2024), that used an interpretable ML framework to directly address the problem of sensor prioritization, demonstrating how XAI can move beyond model explanation to guide strategic decisions in network design and management. Ultimately, the goal is to foster a responsible and effective use of AI in hydrology, where the predictive power of ML is guided and validated by human expertise and domain knowledge (Maier et al., 2024).

3.5 Research Gap

This review has traced the parallel evolution of several key domains in flood risk management: the establishment of comprehensive FEWS, the methodological differences between process-based and data-driven forecasting models, the operational challenge of monitoring network data integrity, and the emergence of XAI to build trust in data-driven models.

The current body of research shows three distinct, but largely separate, streams of development. The first is the ongoing quest for higher predictive accuracy, where complex ML model architectures are consistently demonstrating superior performance. The second is the acknowledgment of operational reality, where studies on network design and data scarcity underscore that real-world monitoring systems are prone to failure. The third stream is the drive for transparency, where XAI methods are being applied to understanding each model’s learning process and identify how the ML models are learning hydrologically relationships.

The literature demonstrates that even though significant research has focused on improving the accuracy of forecasting models, and XAI has been used to understand

their internal logic, a critical research gap exists. Recent work, such as Menapace et al. (2025), has successfully applied interpretable machine learning to create a static prioritization of sensors, providing a valuable framework for maintenance planning. However, this approach primarily addresses the question of which sensors are important in a fully functional network, just removing entire gauging station's data. What remains underexplored is the more dynamic and operationally critical question of how the entire system behaves under different stress conditions. There is opportunity for a framework that uses XAI not just for static ranking, but as a diagnostic tool to systematically assess the resilience of the forecasting system to the inevitable reality of network failures. By integrating these separate research stream, XAI to interpret model performance under progressive data deterioration and in different seasonal conditions, this thesis aims to contribute to the state of art in this gap, providing a novel methodology for assessing and enhancing the resilience of flood monitoring networks.

4 METHODOLOGY

This chapter outlines the methodological approach adopted to evaluate the resilience of flood monitoring networks and prioritize its gauges' importance using an XAI framework. The methodology is structured to first describe the study area (subchapter 4.1) and the dataset adopted in this research (subchapter 4.2), followed by the development and training of the machine learning model (subchapter 4.3). Subsequently, the experimental design for testing network resilience through simulated data failures (subchapter 4.4) and the application of XAI techniques for model interpretation is detailed (subchapter 4.5). Finally, the integrated framework used to synthesize these results for a comprehensive resilience assessment is described (subchapter 4.6).

4.1 Study Area

The Piracicaba River basin is a significant component of the larger Piracicaba-Capivari-Jundiaí (PCJ) river basin system, which spans across the states of São Paulo and Minas Gerais in southeastern Brazil. The basin has critical importance due to its high population density, industrial and agricultural activity, and historical susceptibility to flood events, particularly in its downstream urban centers. Piracicaba City was

selected as the focal point of this study due to its historical vulnerability to inundation and data availability. As characterized by Ferraz et al. (1998), the combination of the city's topography and intensifying urban occupation has established distinct flood-prone zones. As illustrated in Figure 4.1, the study area comprises part of the Piracicaba River basin, which is formed by the confluence of the Atibaia and Jaguari rivers, which are its main tributaries.

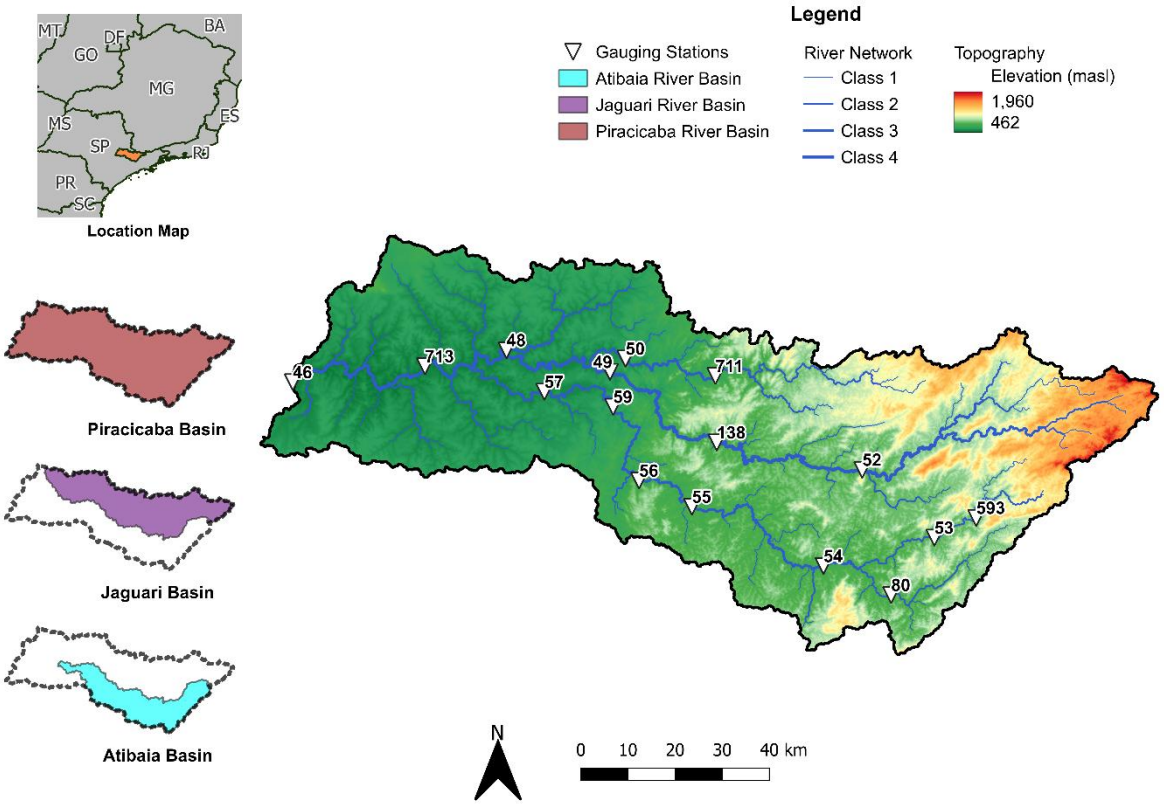


Figure 1 – Location, topography, and hydro-meteorological monitoring network of the Piracicaba River basin

Source: The author (2025).

The topography plays a crucial role in its hydrological response. The headwaters, located in the eastern portion of the basin, are characterized by higher elevations, reaching up to 1,960 meters above sea level (masl). This mountainous terrain contributes to rapid runoff generation during rainfall events. In the downstream portion, the terrain transitions to a gently sloping plain with elevations close to 462 masl, where the main urban areas, including the city of Piracicaba, are located. These topographical gradient influences both runoff onset and the flood wave propagation. In the upstream portion of the basin, Ferrasols are more proeminent, while in the middle and

downstream parts, Lixisols and Rhodic Ferrasols are encountered (Machado et al., 2025). The middle and downstream portions of the basin are also the more urbanized ones, thus contributing more intensely with runoff. Moreover, land use has shifted considerably over recent decades across the basin, marked by the expansion of sugarcane agroindustry and urban areas, often replacing pasture and native vegetation. Projections suggest these trends may continue, potentially altering runoff characteristics in the Piracicaba Basin (Machado et al.,2025). The selected Piracicaba River subbasin has an area of approximately 8889 square kilometers (km²), while the selected Atibaia River subbasin has an area of 2814 km² and the selected Jaguari subbasin has an area of 4340 km².

With a population of 423,323 (IBGE, 2022), this urban center is particularly vulnerable to flooding, making accurate and timely forecasts essential. The basin's concentration time, a key indicator of its response speed, was estimated using the Kirpich equation, resulting in approximately 56 hours. The geographical distribution of these monitoring stations in relation to the river network and topography is shown in Figure 1.

4.2 Monitoring network, data acquisition and pre-processing

This section details the process of acquiring the necessary hydro-meteorological data, the sources from which it was obtained, and the preprocessing steps to prepare it for use in the machine learning model.

4.2.1 Dataset Description and Sources

The Piracicaba River basin is monitored by a hydro-meteorological sensor network composed of gauges that provide both water level and precipitation data. For this study, a network of 16 gauges distributed across the basin was selected to capture its hydrological variability. The target point for the flood forecasting model is gauge 46, located at the downstream end of the basin in the city of Piracicaba.

The gauges were filtered by time-series data availability, using 10% failure threshold to select only the stations with admissible gaps. For each station, two variables were used: water level, measured in meters, and accumulated precipitation, measured in millimeters. The data is available at a high temporal resolution of 10 minutes, which

could be useful for capturing the dynamics of rapidly developing flood events and providing timely forecasts.

The historical data selected spans a continuous five-year period, from January 1, 2018, to December 31, 2022. This timeframe was chosen to include a representative range of hydrological conditions, encompassing multiple wet and dry seasons and flood events of different magnitudes, with minimum gaps within the time-series. The data was sourced from the Department of Waters and Energy (DAEE-SP), the São Paulo agency responsible for managing water resources and operating the state hydrometric network. The general hydrological behavior of the basin during this period is illustrated in Figure 4.2, which shows the rainfall input and the corresponding water level response at the outlet gauge (gauge 46 in Figure 1). This visualization clearly depicts the seasonal patterns and the cause-and-effect relationship between precipitation events and river level fluctuations during the different seasons. Each selected rainfall and water level gauges can be found in Appendix A.

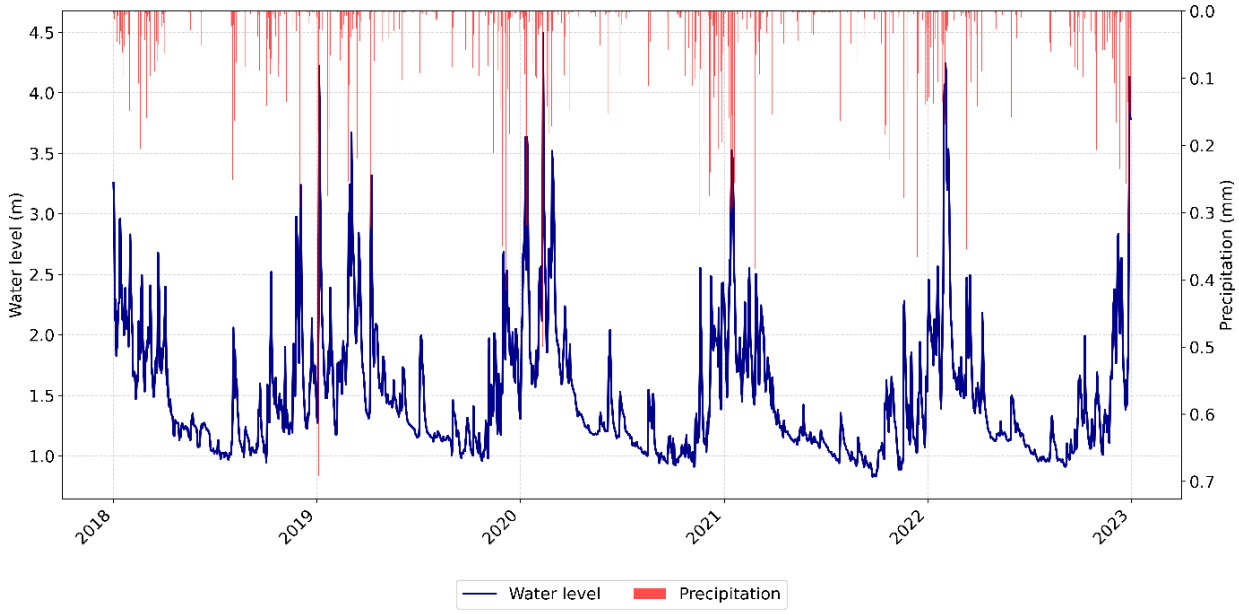


Figure 2 – Hydro-meteorological time-series for the study window (2018-2022) in Gauge 46.
Source: The author (2025).

4.2.2 Data Filling and Partitioning

Field measured data is often subject to imperfections such as missing values, which can arise from sensor malfunctions, power outages, or data transmission errors. As previous stated, the dataset adopted for this research was filtered to allow maximum

10% data gaps in each gauge variable. For the data-driven modeling, it is necessary to ensure the continuity of the time-series, so a data processing protocol was implemented. Missing data points were handled using linear interpolation, a standard method chosen for its simplicity and effectiveness in filling short, incidental gaps without introducing significant bias or distorting the natural variability of the data (Kornelsen & Coulibaly, 2014). Following the filling process, the complete dataset was partitioned into two subsets for model training and validation. Division of the data set into validation and training sets is a crucial step when building a model since it avoids the test being biased by information that the model has previously encountered (Hastie et al., 2009). Splitting the data is done to ensure that the validation accounts for the ability of the model to generalize new data that is yet unknown and is essential in determining the predictive ability and robustness of the model when deployed in real scenarios.

As shown in Figure 3, the first 80% of the dataset (from January 2018 to December 2021) was allocated as the training set, used for the model to learn the complex relationships between rainfall, runoff, and river stage. The final 20% of the dataset (from December 2021 to December 2022) was reserved as the validation dataset. This held-out set was used exclusively for evaluating the final model's performance and for conducting the subsequent data deterioration and resilience experiments, providing an unbiased assessment of the model's generalization capability.

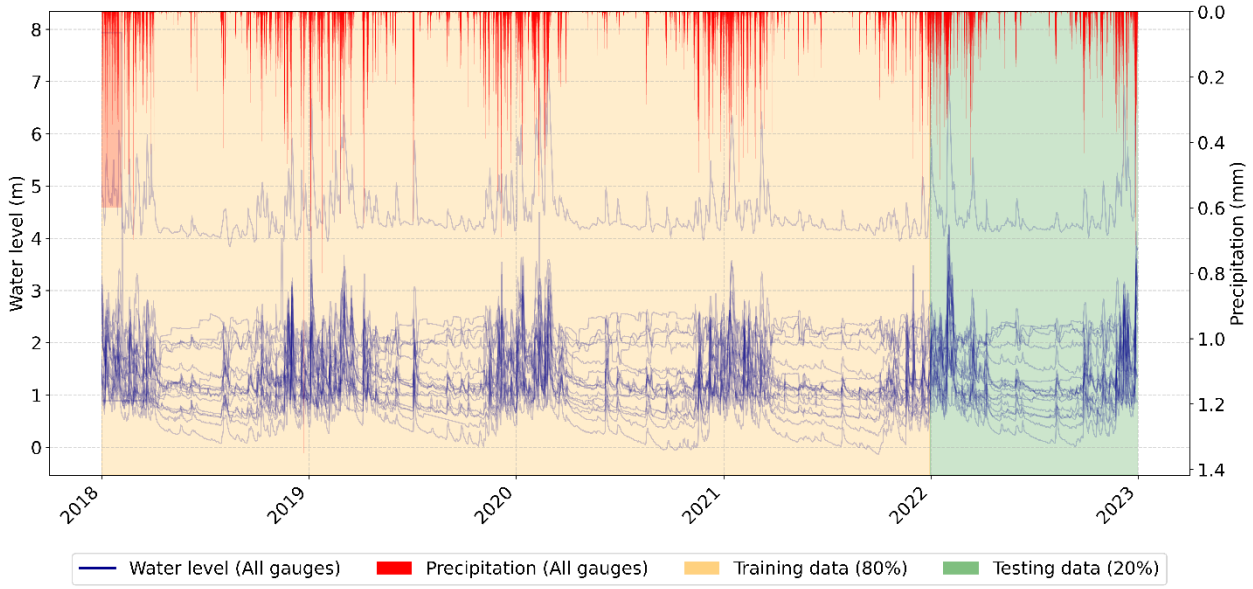


Figure 3 – The 16 hydro-meteorological time-series partition for model training (80%) and validation (20%).

Source: The author (2025).

4.3 Machine Learning Framework

An MLP, a type of feedforward ANN, was selected as the core of the flood forecasting model. The choice was motivated by the MLP's established effectiveness in approximating complex non-linear functions, its suitability for time-series regression problems, and its relative simplicity and computational efficiency (Schalkoff, 1997).

4.3.1 Model Architecture and Input Configuration

The MLP model is structured as a fully connected network consisting of an input layer, one or more hidden layers, and an output layer (Haykin, 2009). For this study, the model receives as input a vector of historical data, which includes past water level and precipitation values from all 16 gauges in the monitoring network. The model's output is a single value: the forecasted water level at the target gauge (46) for a specific lead time.

To address the operational needs of a flood early warning system, which requires forecasts at various lead times, a separate and independent MLP model was trained for each desired lead time: 3, 6, 9, 12, 18, and 24 hours. This approach results in six

specialized models, each optimized for a specific lead time, as recommended by the World Meteorological Organization (2015) for short-term forecasting. The final architecture for each model, determined by grid searching, consists of 31 inputs, a single hidden layer with 128 neurons, and a single-neuron output layer, as presented in Figure 4. This architecture was achieved since each gauge has both rainfall and water-level data and 16 gauges were adopted for the modeling, and the model should predict water level in the target gauge (46). The hidden layer size and number of neurons was defined by a gridsearch.

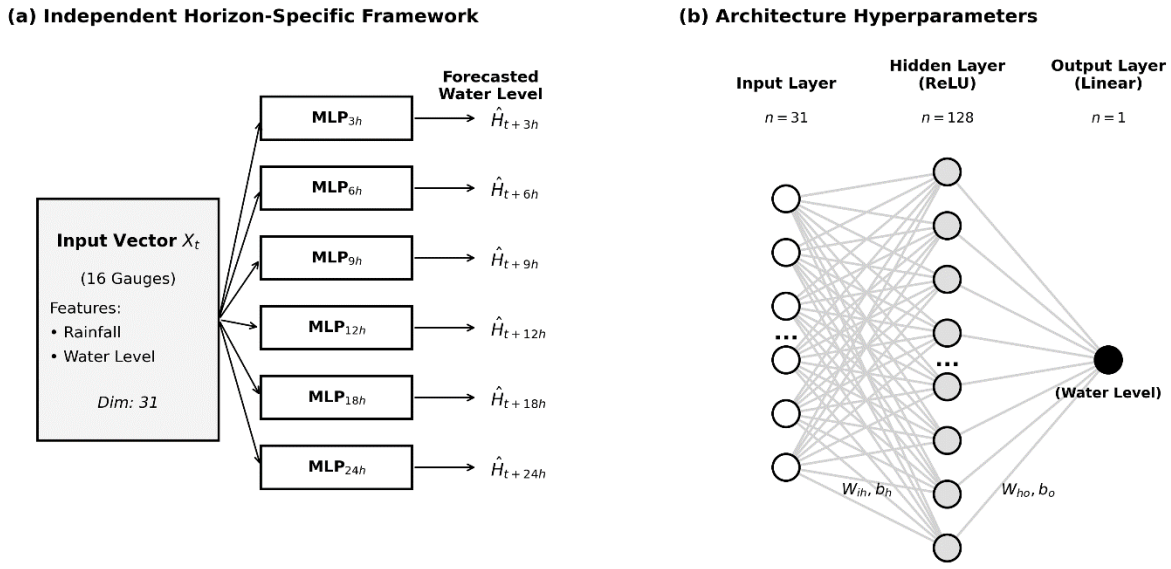


Figure 4 – Adopted MLP Modeling Architechture for different lead times.
Source: The author (2025).

This analysis frames the specific learning task for the MLP model, to capture the complex, non-linear transformation of a sparse, highly variable, and intermittent rainfall input into a continuous, time-lagged, and attenuated water level output, while accurately predicting the extreme peaks that constitute flood events. Recognizing these distinct data characteristics is fundamental not only for model accuracy but also for interpreting its performance and resilience under the stress scenarios explored in subsequent sections.

4.3.2 Mathematical Formulation and Learning Process

The fundamental operation within an MLP is the neuron. During the forward propagation phase, each neuron in a layer receives inputs from all neurons in the

preceding layer. It computes a weighted sum of these inputs, adds a bias term, and then passes the result through a non-linear activation function. Mathematically, the output of a neuron is given by:

$$\hat{y} = f\left(\sum_{i=1}^{n} w_i x_i + b\right) \qquad \text{Equation 1}$$

- x_i are the input values;
- w_i are the weights;
- b is the bias;
- $f(x)$ is the activation function.

The activation function introduces non-linearity into the model, which is crucial for learning complex patterns that cannot be captured by a simple linear model (Hornik et al., 1989; Bishop, 1995; Goodfellow et al., 2016). For this research, the Rectified Linear Unit (ReLU) was chosen as the activation function for the hidden layers, defined as $f(x) = \max(0, x)$. The ReLU function is widely used due to its computational efficiency and its effectiveness in mitigating the vanishing gradient problem during training (Glorot and Bengio, 2010).

The network calibrates its internal parameters through a process called backpropagation (Rumelhart et al., 1986, Haykin, 2009). After a forward pass, the model's prediction is compared to the actual observed value aiming to optimize a loss function. For this regression task, the Mean Squared Error (MSE) was used, which measures the average squared difference between the observed and predicted values:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad \text{Equation 2}$$

- n is the number of observations;
- y_i is the observed value in the time step i;
- $\hat{y}_i$ is the forecasted value in the time step i;

The backpropagation algorithm calculates the gradient of this loss function with respect to each weight and bias in the network. These gradients are then used by an optimization algorithm to update the weights and biases in the direction that minimizes the loss. For this study, the Adaptive Moment Estimation (Adam) optimizer was

employed, a sophisticated algorithm that adapts the learning rate for each parameter, leading to efficient convergence (Kingma & Ba, 2014). This iterative process of forward propagation, loss calculation, and backpropagation is repeated for 75 epochs until the model's performance on a validation set ceases to improve, thus preventing overfitting. Figure 5 shows a schematic for the ML framework adopted.

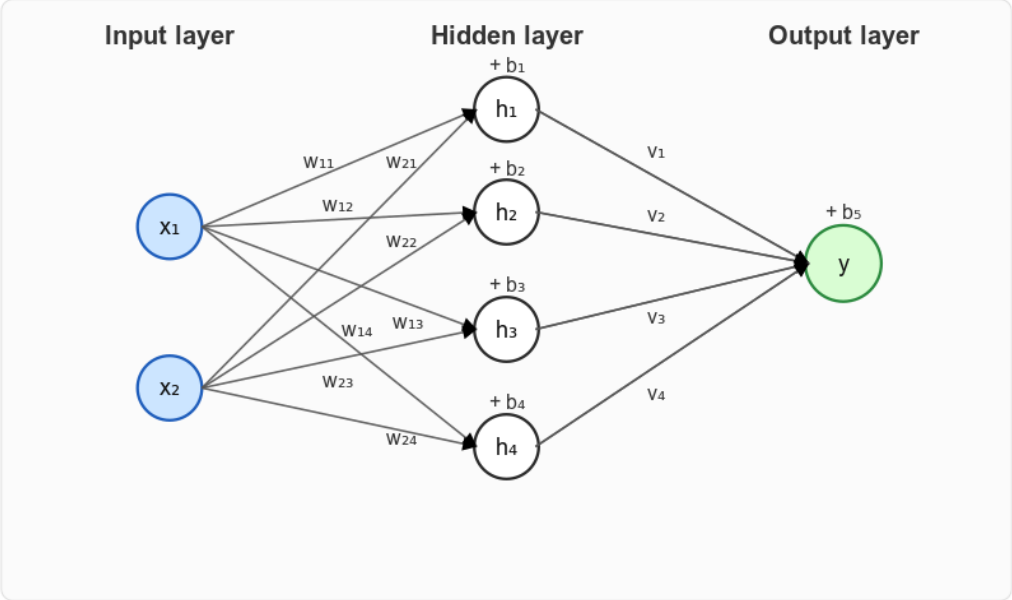


Figure 5 – MLP Schematic for a 2 input layers, 4 neurons and 1 output layer.
Source: The author (2025).

4.3.3 Performance Evaluation

While MSE is used to train the model, its performance on the unseen test set was primarily evaluated using the NSE coefficient (Nash & Sutcliffe, 1970) and the KGE coefficient (Gupta et al., 2009).

The NSE is a normalized statistic that determines the relative magnitude of the residual variance compared to the observed data variance. It is a standard performance metric in hydrology and is calculated as:

$$NSE = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad \text{Equation 3}$$

- n is the number of observations;
- y_i is the observed value in the time step i;

- $\hat{y}_i$ is the forecasted value in the time step i;
- $\bar{y}$ is the mean of observed time-series.

An NSE of 1 corresponds to a perfect match between predicted and observed data. An NSE of 0 indicates that the model predictions are as accurate as the mean of the observed data, while an NSE less than 0 suggests that the observed mean is a better predictor than the model.

The Kling–Gupta Efficiency (KGE) was also calculated to provide a different approach to performance evaluation by decomposing model errors into correlation, bias, and variability components. The KGE is defined and calculated by:

$$KGE = 1 - \sqrt{(r-1)^2 + (\beta-1)^2 + (\gamma-1)^2} \quad \text{Equation 4}$$

- r is the Pearson correlation coefficient between observed and simulated data;
- $\beta = \frac{\mu_s}{\mu_0}$ is the bias ratio, comparing the mean of simulated (μ_s) and observed data (μ_0);
- $\gamma = \frac{CV_s}{CV_0} = \frac{\sigma_s/\mu_s}{\sigma_0/\mu_0}$ is the variability ratio, comparing the coefficients of variation of simulated and observed series.

A KGE value of 1 indicates perfect fit between observed and simulated data, while values close to zero indicates less accurate performances. Negative KGE values indicate that the model performs worse than a simple mean-based prediction.

4.4 Experimental Design for Network Resilience Analysis

To quantify the resilience of the monitoring network and identify critical gauges, a systematic data deterioration experiment was designed. This involved intentionally degrading the quality of the input data fed to the already-trained MLP models and observing the impact on forecast accuracy.

4.4.1 The Time-Series Deterioration Algorithm

A custom algorithm was developed to simulate sensor failure in a realistic manner. This algorithm is based on the methodology proposed by Menapace et al. (2025), which they validated in the South Tyrol basin. However, instead of removing individual data points, which is less representative of a real-world outage, the proposed refined algorithm applies deterioration in continuous blocks. For a selected gauge, the

algorithm randomly selects starting points within the validation set's time-series. From each starting point, it replaces a block of consecutive data points with zeros, simulating a complete loss of signal. The length of each block is set to be equal to the forecasting lead time of the model being tested (e.g., for the 3-hour forecast model, the block length is 3 hours of 10-minute intervals). Extending the approach presented by Menapace et al. (2025), this study implements a progressive data deterioration, ranging from 10% to 100% of each gauge's time series. This analysis aims to identify the sensors resilience thresholds, quantifying the extent of data loss each gauge can sustain before forecast accuracy is compromised.

The algorithm was applied only to the water level (FLU) data for each gauge in an iterative manner, as this variable is more directly indicative of a gauge's operational status for flood routing and has a stronger autocorrelation, making its loss more impactful than intermittent rainfall data (Menapace et al. 2025). To ensure the experiment is repeatable, a fixed random seed equal to 12 was used for the selection of deterioration blocks. An example of deteriorated time series, for gauge 713, is illustrated in Figure 6.

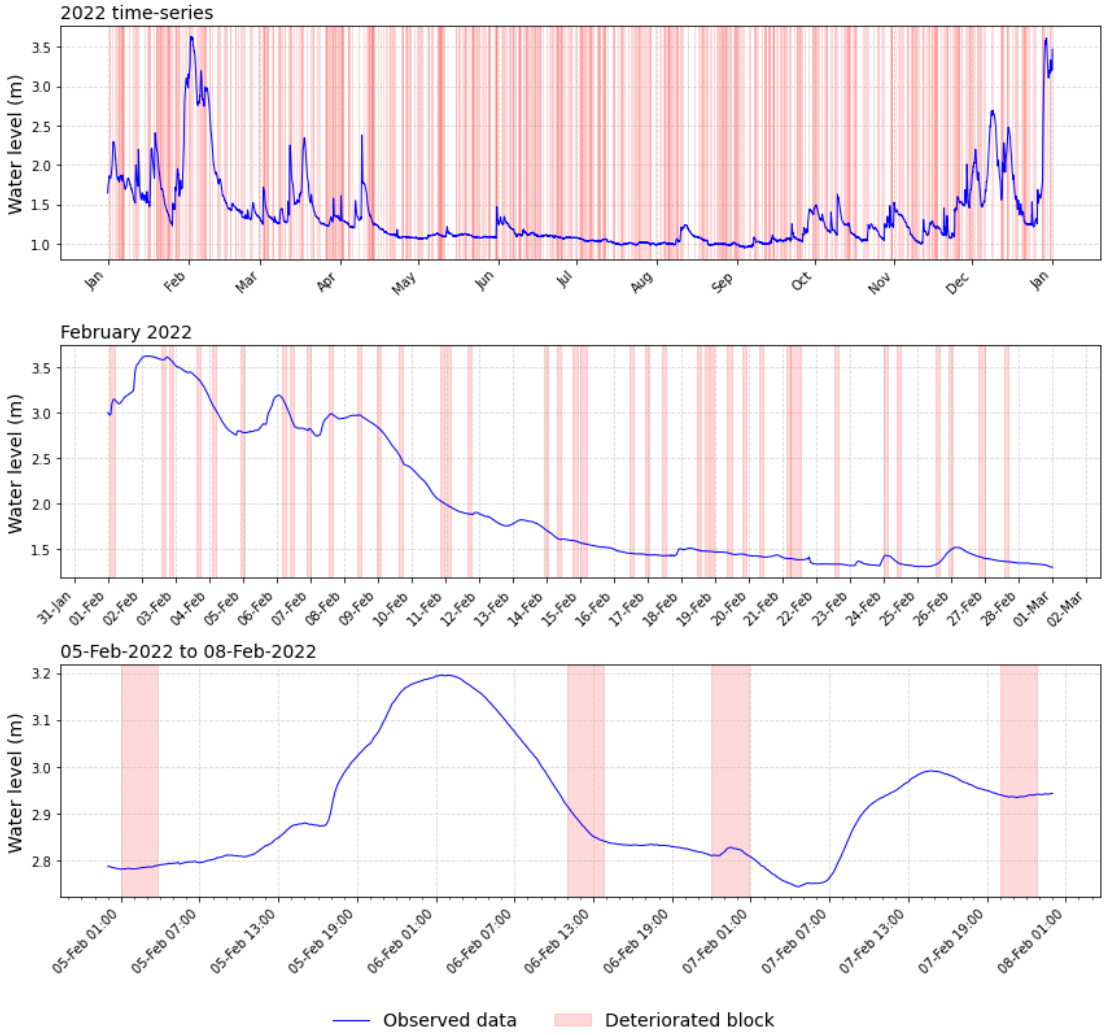


Figure 6 – Time series deterioration in blocks in gauge 713 for 20% data loss.
Source: The author (2025).

4.4.2 Scenario Definition

The deterioration process was applied under a comprehensive set of scenarios designed to systematically probe the network's weaknesses. This multi-faceted approach allows for a robust analysis of each gauge's importance under various conditions. The scenarios are defined by three key dimensions: the specific gauge being targeted, the percentage of data being removed, and the seasonal period of the failure.

The core of the experiment involved isolating the impact of individual sensor failures. The deterioration algorithm was applied independently to each of the 15 upstream monitoring stations, one at a time, while keeping the data from all other stations intact. This "one-at-a-time" sensitivity analysis is crucial for attributing any drop in model

performance directly to the loss of information from a single, specific gauge. This allows for a clear ranking of the stations based on their individual impact on forecast accuracy.

Subsequently, to understand the relationship between the extent of data loss and forecast accuracy, the deterioration for each gauge was applied at progressive levels of severity. Specifically, scenarios were run for 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, and 100% data loss. A 0% scenario was also included as a baseline. This progressive failure analysis is designed to reveal the sensitivity of the forecasting system to varying degrees of sensor malfunction. It helps to identify whether the impact of a gauge is linear (i.e., a 20% loss is twice as bad as a 10% loss) or if there are non-linear effects and critical thresholds beyond which forecast quality collapses abruptly. To quantify the operational resilience of the forecasting system on the progressive data failure in each gauge, three performance thresholds were defined based on the NSE: Optimal (NSE ≥ 0.90), representing ideal forecasting conditions, degraded (0.75 ≤ NSE < 0.90), where forecasts remain feasible but with reduced confidence, and Critical Failure (NSE ≤ 0.50), indicating the model is no longer reliable for flood warning. Model interpretations when predictive performance exceeded an NSE of 0.70, in line with Bayati et al. (2025). The KGE metric was adopted to further quantify the operational resilience, visualized through heatmaps that illustrate the gradient of performance decay.

Finally, regarding seasonality, the hydrological response of the Piracicaba basin is highly seasonal, with a distinct wet season (October to March) characterized by convective storms and high flood risk, and a dry season (April to September) with lower flows. To investigate whether a gauge's importance varies between these periods, the deterioration was applied selectively based on season. For each gauge and percentage, three separate tests were conducted: one where data was deteriorated only during the wet season months, one where it was deteriorated only during the dry season, and one where the deterioration was applied across the entire year regardless of season. This provides crucial insights into which gauges are most critical for forecasting during the primary flood season, which is of greatest concern for operational warning systems.

The framework proposed, encompassing hundreds of individual scenarios (15 stations × 10 percentages × 3 seasonal options), allows mapping of how forecast accuracy, measured by the drop in NSE, is affected by the loss of information from each specific gauge, at different levels of failure, and across the two different seasons. Figure 7 shows the deterioration in blocks per season, as described in this subchapter.

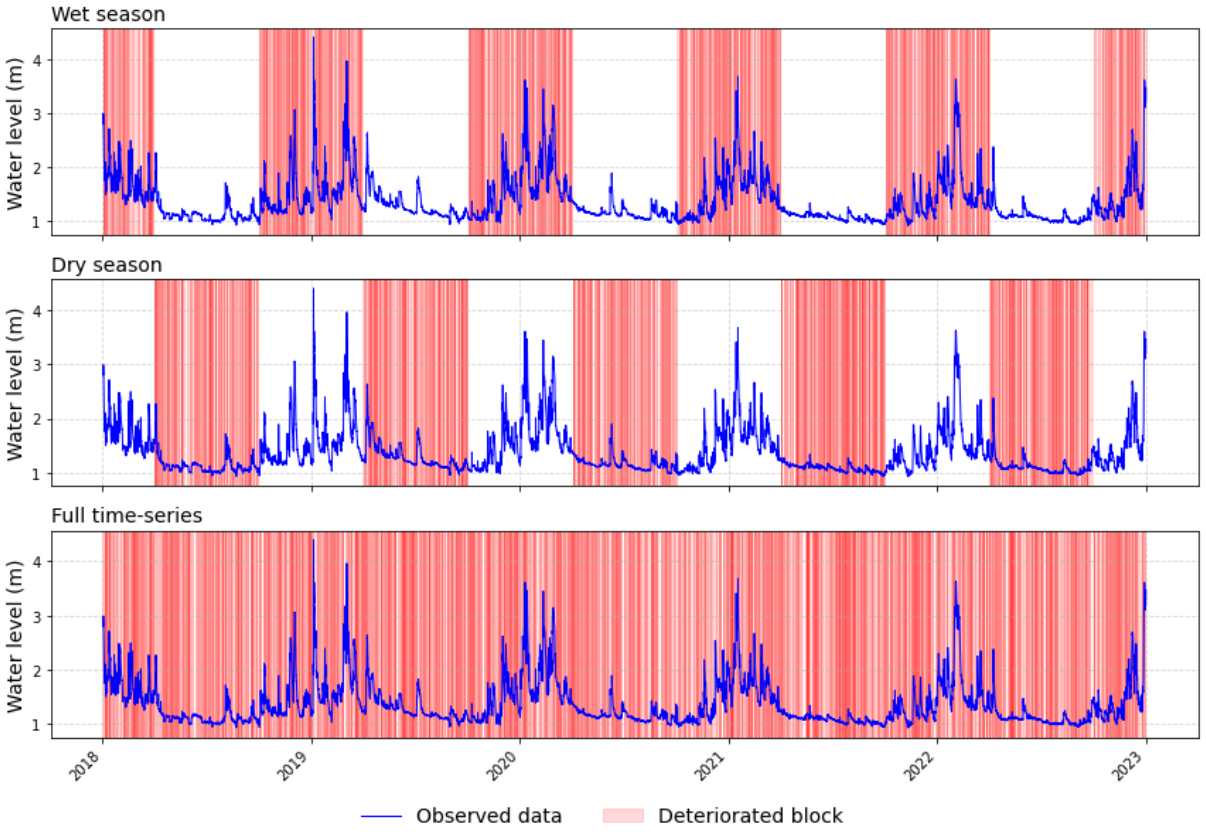


Figure 7 – Time series deterioration in blocks in gauge 713 for 20% data loss (seasonal).
Source: The author (2025).

4.5 Interpretation and Gauge Prioritization with XAI

To move beyond simply observing model performance and to assess if gauges importance for FEWS resilience is hydrologically reasonable, an XAI technique was integrated into the analysis. This step is crucial for building trust in the model's predictions and for validating that its internal logic aligns with established hydrological principles.

4.5.1 SHAP for Quantifying Gauge Contribution

SHAP (SHapley Additive exPlanations) was selected as the XAI method for this research (Lundberg & Lee, 2017). SHAP is a state-of-the-art technique rooted in cooperative game theory that provides a rigorous and theoretically sound approach to explaining the output of any machine learning model (Lundberg & Lee, 2017). SHAP

calculates the fairest way to distribute the difference between the model's actual prediction and a baseline average prediction among these features.

For every individual prediction the model makes, each feature is assigned a SHAP value. The magnitude of this value represents the feature's importance or contribution strength, while the sign indicates the direction of the influence. A positive SHAP value signifies that the feature pushes the prediction higher than the baseline (high upstream levels increasing the forecast), whereas a negative value pushes it lower. A key property of SHAP is its additivity; the sum of the SHAP values for all features, when added to the baseline, precisely equals the model's final output (Lundberg & Lee, 2017).

4.5.2 Application of SHAP to trained models

For each lead time, the SHAP methodology was applied to each one of the six MLP - based FEWS. Using the SHAP Python library, importance values were computed for the validation set. This was done to derive a global understanding of feature importance, aggregating the insights from thousands of individual predictions to reveal overarching patterns in the model's behavior. To initialize the SHAP explainer, a representative background dataset is required to serve as a baseline for feature attribution. For this purpose, a flood event from the training period (January 2018) was adopted as the reference sample, shown in Figure 8. A 24-hour window was specifically selected to align with the model's maximum forecasting horizon. This specific event was selected because it captures the complete hydrological morphology of a flood wave, encompassing the hydrograph’s rising, peak, and recession, ensuring the explainer accurately attributes importance during both low and high-flow conditions.

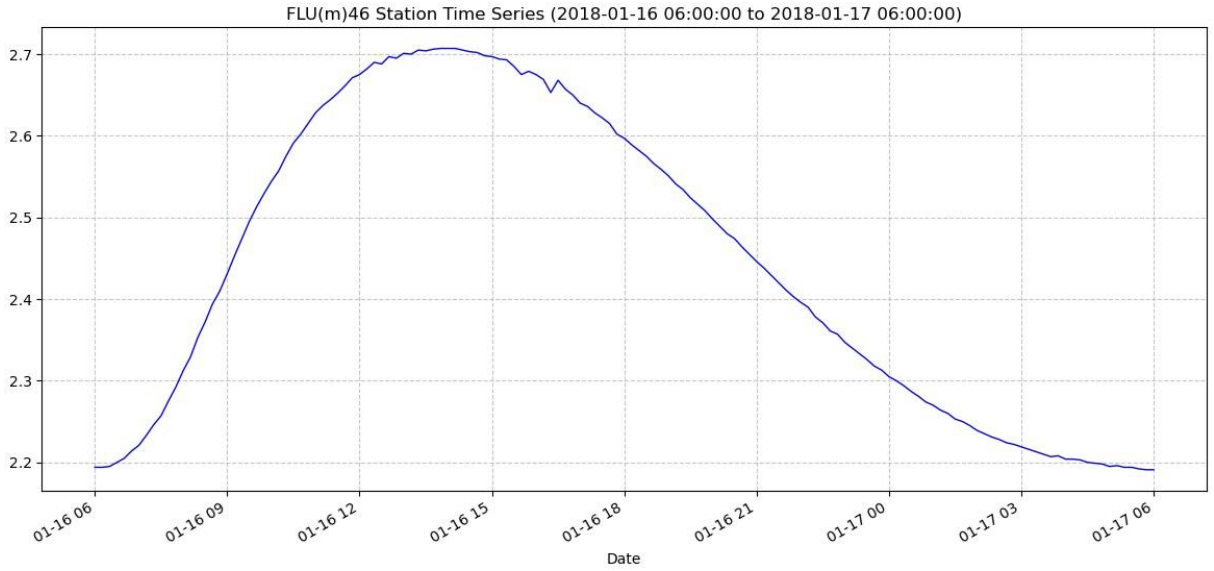


Figure 8 – Predicted Hydrograph used for the SHAP Explainer
Source: The author (2025).

The primary outputs of this analysis are the magnitude and direction of each gauge's influence on water level forecasting in the different lead times. This SHAP plot ranks all gauges by their overall importance and shows the distribution of SHAP values for each, revealing not just which gauges are important, but also how they influence predictions. Additionally, the mean absolute SHAP value for each gauge was calculated for each lead time. This provides a clear, quantitative metric for ranking gauge importance, which can be compiled into a summary table to track how a gauge's influence changes as the forecast lead time increases.

4.6 Integrated Framework for Resilience Assessment

The final and most critical stage of the methodology is the synthesis of the results from the two preceding analytical streams: data deterioration experiments and XAI analysis. This integrated framework is designed to move beyond individual analyses to a holistic assessment of network resilience, providing a robust, multi-faceted approach to prioritizing monitoring gauges, focusing on both data consistency and hydrological representation for assertive and reliable forecasting in data-driven FEWS applications. The integrated framework also enables to validate the physical integrity of the failure methodology proposed by Menapace et al. (2025).

The core of the framework is a direct integration of the gauge rankings produced by each method. First, from the deterioration analysis, a ranking is established based on the magnitude of the drop in the NSE metric. Gauges whose failure causes the most significant decrease in forecast accuracy are ranked as most critical. Then, from the XAI analysis, a second ranking is established based on the mean absolute SHAP values. Gauges that the model consistently relies on most heavily to make its predictions (i.e., those with the highest SHAP values) are ranked as most important.

The deterioration analysis serves as an operational stress test, simulating sensor failure to quantify the system's vulnerability to data loss. Complementarily, the XAI analysis exposes the internal logic of the 'black box' model, indicating which features drive the forecasting logic. By synthesizing these perspectives, the framework moves beyond simple comparison to a broader integration, ensuring that the final gauge prioritization reflects both the mathematical reasoning of the AI and the operational resilience required for real-world flood monitoring. Figure 9 provides a schematic overview of the proposed methodological framework.

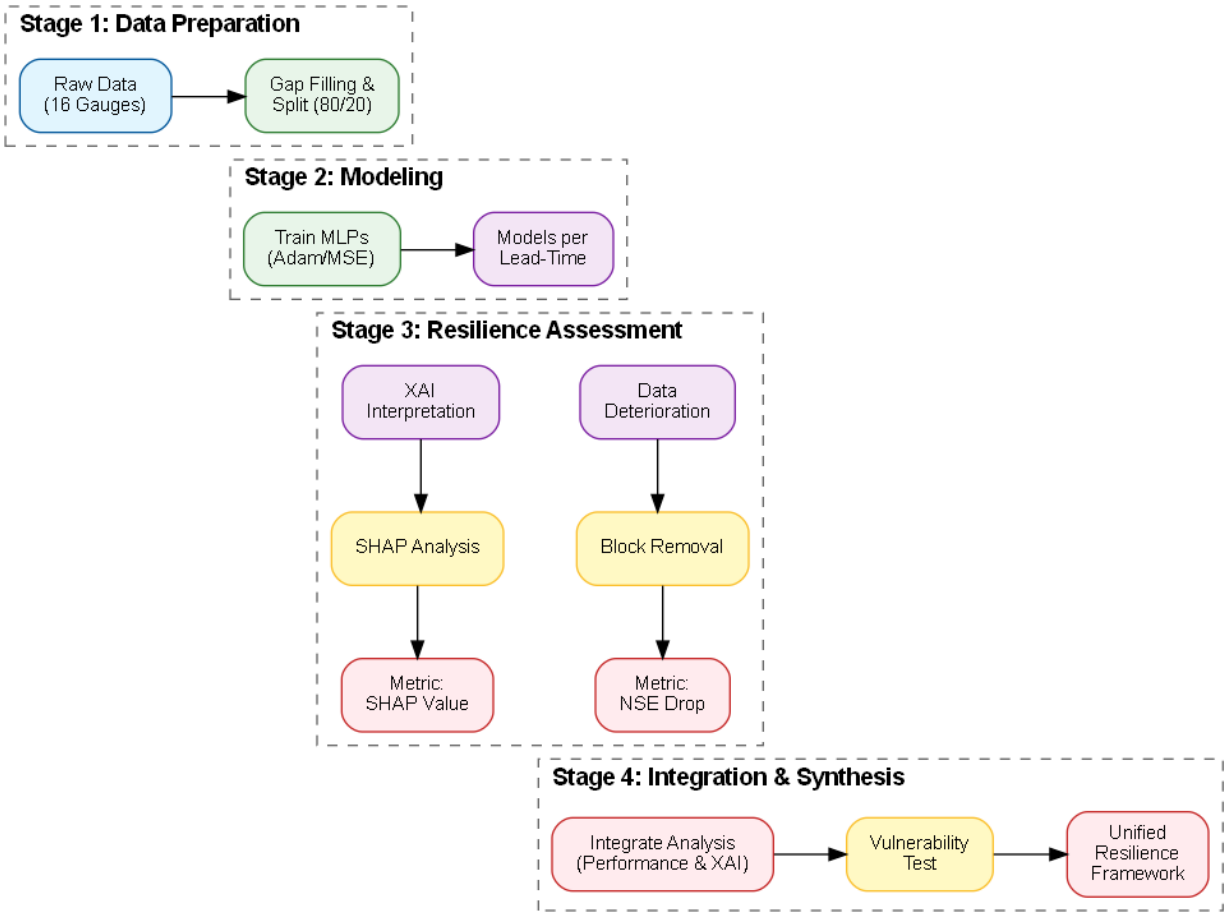


Figure 9 – Integrated Framework for Resilience Assessment flowchart.
Source: The author (2025).

A strong correlation between these two independent rankings provides powerful, cross-validated evidence for the importance of a gauge. For example, if a gauge is ranked highly by both methods, it means that the model relies heavily on its data and that the system is highly vulnerable to its failure. This dual confirmation allows for a confident prioritization of gauges for maintenance and operational attention. Moreover, this provides information to relate the hydrological processes behind such highlighted importance, enabling the advancement of runoff onset and flood wave routing processes in the forecasting system.

As a final step to test the practical implications of this framework, a vulnerability analysis was conducted. A new, "optimized" MLP model was trained from scratch, but this time using only the data from the top five most critical gauges identified by the integrated framework. This lean model was then subjected to the same rigorous deterioration analysis as the full model. The purpose of this test is twofold: first, to assess the performance of a potentially more cost-effective, redundancy-reduced network (Chen et al., 2023), and second, to determine how brittle this optimized network is. By observing how quickly its performance degrades when one of its few critical sensors fails, this analysis provides crucial insights into the trade-offs between network efficiency and resilience, directly addressing the value of redundancy (Jung & Axelsson, 2024). in maintaining a accurate early warning system.

5 RESULTS

This chapter presents the direct findings obtained from the application of the methodology described in Chapter 4. The focus of this section is on the objective presentation of data and experimental outcomes, laying the factual groundwork for the interpretive discussion that will follow in Chapter 6. The results are organized sequentially, beginning with a detailed analysis of the dataset's characteristics, followed by the baseline performance of the forecasting models. Subsequently, the quantitative impacts of simulated data failures are presented, followed by the insights derived from the eXplainable AI analysis. The chapter concludes by presenting the results of the vulnerability analysis performed on a reduced, prioritized network.

5.1 Dataset Analysis and Hydrological Patterns

As stated in the literature review, a good quality dataset is indispensable for accurate forecasts in data-driven models. The 5-year time-series dataset comprises 262,944 10-minute observations for each of the 16 hydrometeriological monitoring stations (rainfall and water level observed data). The descriptive statistical analysis is presented in Table 1 for the water level data-set.

Table 1 – Descriptive statistics for the water level data-set.

Gauge	Mean (μ)	Standard deviaton (σ)	Minimum	25%	50%	75%	Maximum
46	1.49	0.543	0.811	1.12	1.30	1.72	5.04
48	0.81	0.504	0.212	0.484	0.623	0.952	4.36
49	1.35	0.349	0.554	1.13	1.22	1.45	4.31
50	0.493	0.495	0.143	0.181	0.357	0.638	4.53
52	1.08	0.178	0.779	0.983	1.04	1.13	2.51
53	2.08	0.417	1.03	1.81	2.26	2.37	5.71
54	2.20	0.339	1.48	1.99	2.16	2.30	3.72
55	4.47	0.499	3.80	4.20	4.29	4.56	8.22
56	1.14	0.337	0.759	0.95	1.01	1.21	3.67
57	2.12	0.269	1.72	1.95	2.03	2.23	4.05
59	0.839	0.328	0.363	0.644	0.730	0.927	4.56
80	1.88	0.397	0.812	1.61	2.00	2.19	2.78
138	1.38	0.318	0.966	1.17	1.27	1.51	3.18
593	0.81	0.260	0.511	0.634	0.732	0.902	3.39
711	1.73	0.302	1.10	1.53	1.67	1.86	4.64
713	1.36	0.413	0.925	1.09	1.20	1.50	4.40

Source: The author (2025).

The rainfall data across all stations exhibit a pattern characteristic of Southeastern Brazil's climate,with intermitent storms through the year. To analyze the magnitude of these events, descriptive statistics were calculated for rain periods only (accumulations > 0 mm), excluding rainless intervals that otherwise dominate the 10 minute resolution dataset.he 25th, 50th (median), and 75th percentiles for 10-minute rainfall accumulations are uniformly 0.0 mm for every gauge. This statistically indicates that over 75% of the observed intervals recorded no precipitation, reflecting the distinct dry periods and the nature of rainfall occurring mainly in the wet season. This sparsity contrasts with the high intensity observed during rainfall events. Maximum 10-minute accumulations show significant spatial variability, ranging from 17.0 mm (Gauge 46) up to 44.2 mm (Gauge 711). When considering 30-minute accumulations, the mean rainfall depth across the gauges ranges from approximately 1.0 mm to 1.6 mm. The

95th percentile reaches up to 6.25 mm at Gauge 711, while maximum depths exceed 40 mm at multiple gauges (59.2 mm at Gauge 711 and 53 mm at Gauge 57). This variability is also observed at the hourly scale, where the mean accumulation during wet hours is approximately 2.0 mm, yet extreme storms generate peaks up to 103.4 mm (Gauge 711) and 67 mm (Gauge 713). At the daily scale (24 hours), the basin exhibits substantial spatial heterogeneity in total accumulation, with maximum observed values ranging from 71.6 mm (Gauge 52) to an extreme of 194.4 mm (Gauge 711).

In contrast to the intermittent rainfall, the water level data display continuity, reflecting the river system's natural integration and storage processes. The statistics show a smooth progression of non-zero values from minimums (e.g., 0.811 m at Gauge 46) through the 25th, 50th, and 75th percentiles. Nevertheless, the water level exhibits a wide dynamic range, indicative of the basin's susceptibility to flooding. At the target gauge (46) in Piracicaba city, for instance, while the water level is below 1.72 m for 75% of the time, the recorded maximum reached 5.04 m during a flood event. This substantial difference between typical levels and extreme peaks highlights the high-impact, seasonal nature of floods in the basin (Neto et al., 2016). Accurately capturing these infrequent but critical high-water events is the paramount objective for the forecasting model within an operational FEWS context.

The relationship between upstream inputs and downstream response is visually exemplified by hydrograph propagation dynamics within the basin. Figure 10 illustrates this process for a specific flood event in January 2018, comparing the water level time series at the target gauge (46) with the station immediately upstream (713). The plot shows the time lag between the flood peaks observed at the immediately upstream station and their subsequent arrival at the downstream target location. It is possible to identify the propagation of the flood wave between the observations in each gauge.

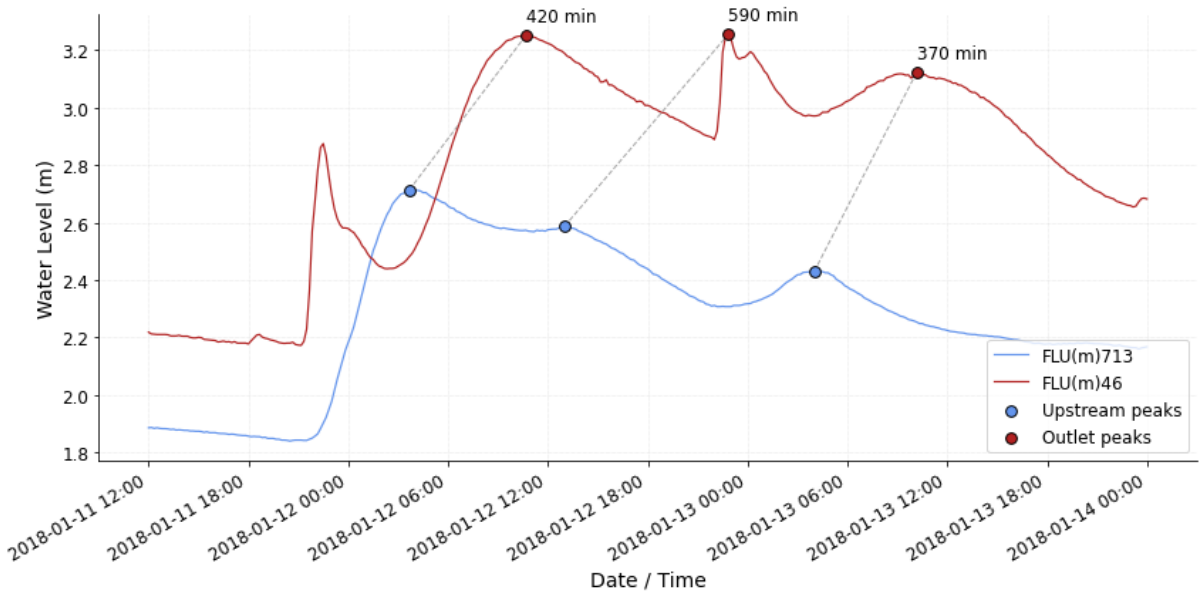


Figure 10 – Hydrograph propagation event in January 2018, between stations 46 (target) and 713 (immediately upstream).

Source: The author (2025).

Figure 11 illustrates this process for a different flood event, in January 2021, comparing the water level time series at the target gauge (46) with a station located in the middle of the basin, upstream (57). With this example, it is noteworthy that a more upstream gauge doesn't have a clear correlation between events as a closer one, but still some events match.

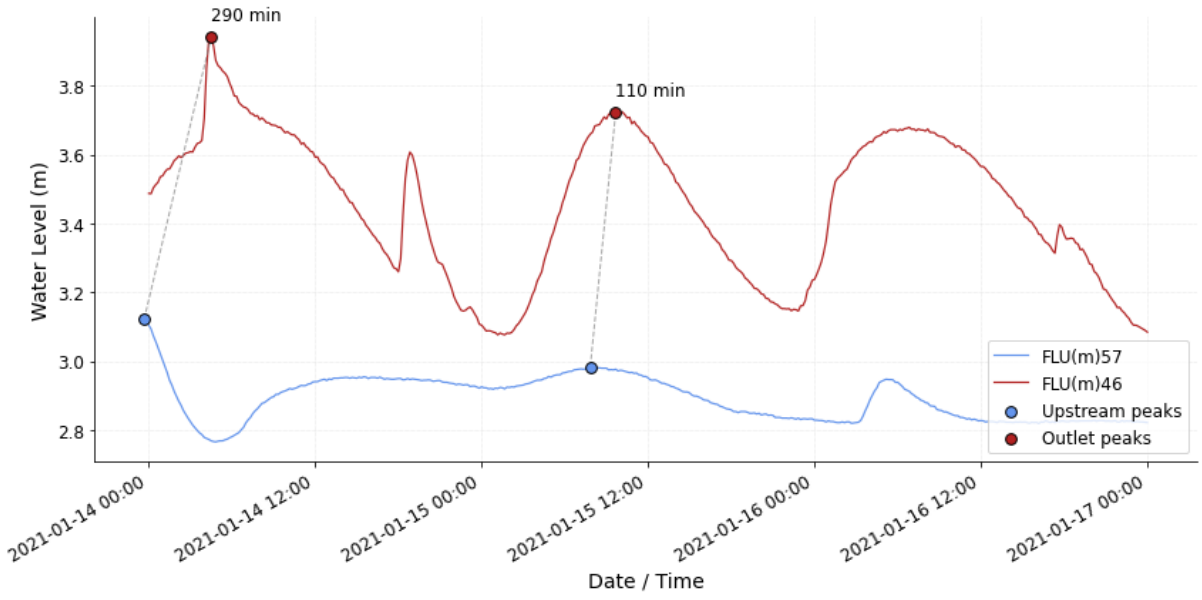


Figure 11 – Hydrograph propagation event in January 2021, between stations 46 (target) and 57 (middle-basin upstream).
Source: The author (2025).

This observed propagation presented in both examples is a fundamental hydrological process governed by channel characteristics and flow dynamics. Capturing this time-dependent relationship accurately is an important task for the forecasting model, as it directly influences the prediction of flood peak timing and magnitude at the target location for the different lead times.

5.2 Baseline Model Performance

Before assessing the network's resilience to data failures, it was first necessary to establish the baseline performance of the MLP forecasting models under complete time series data. The performance of the six models, each trained for a different lead time, was evaluated using the NSE and KGE coefficients.

The results indicate the model accuracy for every lead time (Table 2). The models for shorter lead times (3 to 12 hours) demonstrate better performance, with NSE values exceeding 0,940 and KGE values above 0,920. As the lead time increases to 18 and 24 hours, a gradual and expected decrease in performance is observed, with both the NSE and KGE values dropping by approximately 5% from the 12-hour to the 24-hour model. Despite this, the performance remains high (NSE > 0,880 and KGE > 0,870), indicating that the MLP architecture could learn rainfall-runoff relationships from the training data and is capable of forecasting across different lead times when provided with good quality input data.

Table 2 – Baseline performance of the MLP models for each lead time.

Lead time (hours)	Hidden size	Hidden layers	NSE	KGE
3	128	1	0,964	0,941
6	128	1	0,967	0,927
9	128	1	0,960	0,929
12	128	1	0,948	0,926
18	128	1	0,913	0,904
24	128	1	0,882	0,879

Source: The author (2025).

To provide a qualitative and quantitative assessment of the model's performance, a hydrograph plot and a scatter plot are presented for both 3-hour and 24-hour lead

times. Figure 12 compares the predicted and observed water level hydrographs over the entire test set. The plot shows a close correspondence between the model's predictions and the observed water level, indicating that the model successfully captures not only the magnitude of the flood peaks but also their timing and the overall shape of the hydrograph. Figure 13 also compares the predicted series for a 24-hour lead time, indicating a high accuracy in spite of the earlier forecast.

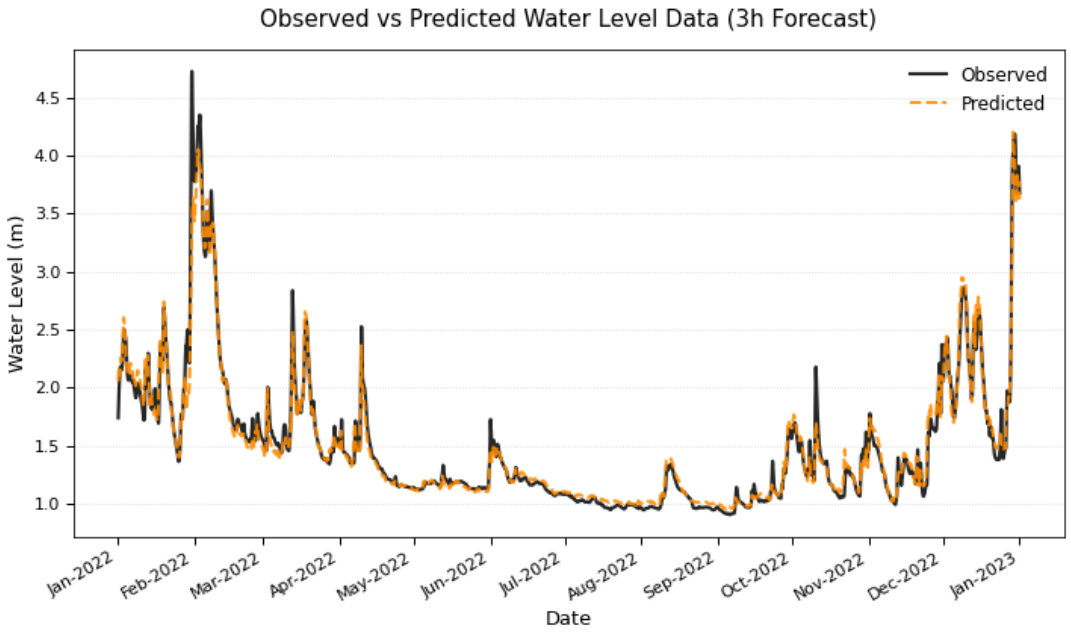


Figure 12 – Water level forecast for 3 hours lead time with the complete time-series input.
Source: The author (2025).

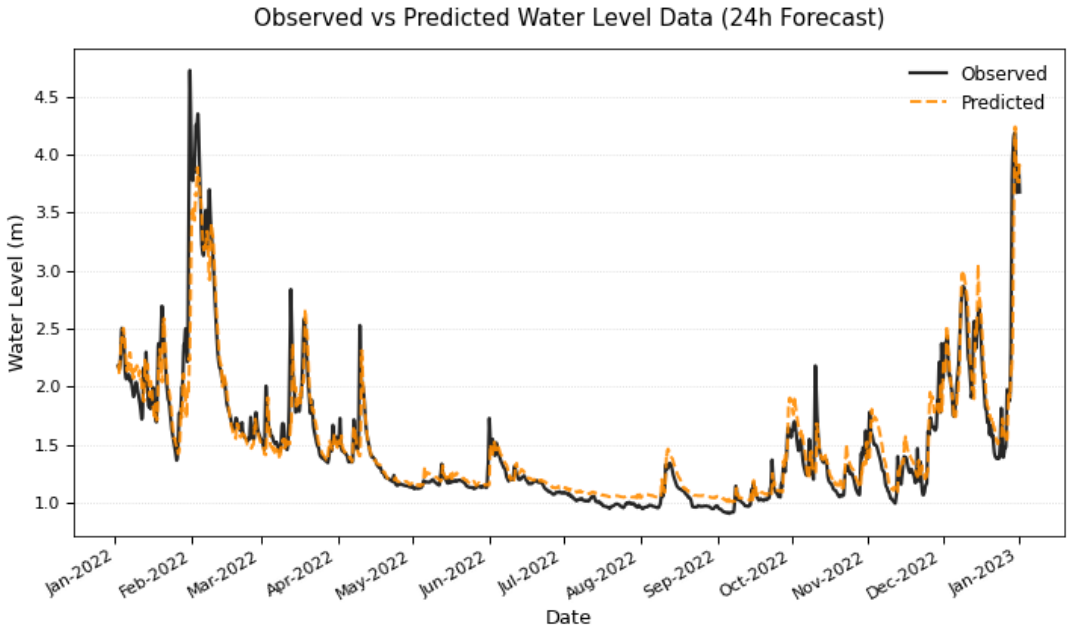


Figure 13 – Water level forecast for 24 hours lead time with the complete time-series input.
Source: The author (2025).

Figure 14A provides a complementary view, showing a scatter plot of the same predicted versus observed values. The high density of points clustered tightly around the diagonal line visually confirms the model's low error, while the further points represents the bias of the model. The high coefficient of determination ($R^2 > 0.8$) further quantifies this strong linear relationship between the predictions and observations. Figure 14B shows a sparser plot, which is an expected result due to the longer lead time.

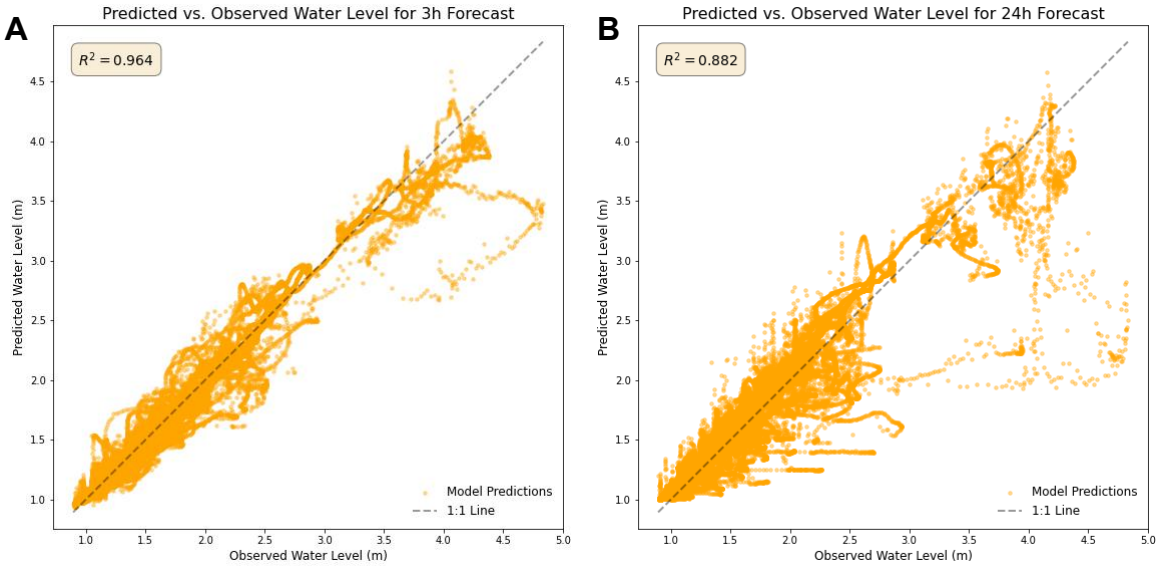


Figure 14 – Scatter plot of predicted vs. observed water levels for 3-hour and 24-hour lead times with complete data-series input.

Source: The author (2025).

5.3 The Impact of Data Failures on Forecast Accuracy

The proposed methodology systematically degraded the input data from each upstream gauge to measure the resulting drop in forecast accuracy. Figure 15 shows the NSE as a function of the deterioration percentage scenarios from 3 hours to 24 hours lead times, for both seasons. Likewise, a heatmap for the KGE and the deterioration percentage scenarios for the 6 lead times and both seasons deterioration in presented in Figure 16. Plots found in Appendix B present the seasonal (wet and dry seasons) deterioration of NSE trends and KGE heatmaps.

For both seasons scenario, the short-term (3–6 hours) system exhibits its fragility regarding the immediate upstream point, Gauge 713. In the 3-hour forecast, the loss of around 20% of data from this sensor causes the model to drop below the 0.75threshold. The system reaches the critical failure point (NSE ≤ 0.50) when data loss is around 35%, rendering the nowcasting inaccurate. This sensitivity persists at the 6-hour horizon, where Gauge 713 remains the primary failure point, crossing the critical threshold at approximately 40% data loss. As the forecast horizon extends to the intermediate range (9–12 Hours), the network's dependency shifts slightly. At 9

and 12 hours, the influence of Gauge 713 reduces, while one of the mid-basin sensors, Gauge 55, begins to contribute to model's performance. The 9-hour forecast starts to become vulnerable to sensors located in the middle of the basin, when Gauge 55 starts to play an important role in degradation accuracy, with 25% failure dropping the model's performance below 0.75. For the 12-hour forecast, while Gauge 713 still degrades performance, the shift towards Gauge 55 starts to become more evident. That sensor data degradation drives the system out of the optimal zone (NSE ≥ 0.7) with less than 10% data loss, marking the transition of the model's 'attention' from the basin outlet to upstream boundary conditions. For the longest lead times (18–24 Hours), the network enters a state of dual dependency. Unlike the short-term scenarios dominated solely by the immediately upstream gauge (Gauge 713), the long-term forecasts rely critically on both Gauge 713 and mid-basin sensor Gauge 55. At the 18-hour horizon, the system remains vulnerable to Gauge 713, which drives performance below the critical threshold when data loss exceeds 80%. The influence of Gauge 55 intensifies, causing the model to reach the degraded threshold at only 25% data loss. For the 24-hour forecast, this shift towards the mid-basin is more pronounced. While the failure of Gauge 713 still compromises the system, dropping below 0.50 at ~80% loss, Gauge 55 becomes the primary driver of early degradation. A loss of just 15% of data from Gauge 55 drops the system into degraded status, and it reaches the critical threshold at 60% loss, slightly earlier than Gauge 713. In this lead-time range, a third sensor also plays a role in the prediction, Gauge 53. For 18-hour forecast, a 100% data deterioration in this sensor makes the model drop below the degraded threshold. In the 24-hour lead time, 70% data loss compromises the forecasting to the degraded condition. This indicates that for longer forecasting, the model essentially requires a chain of information:

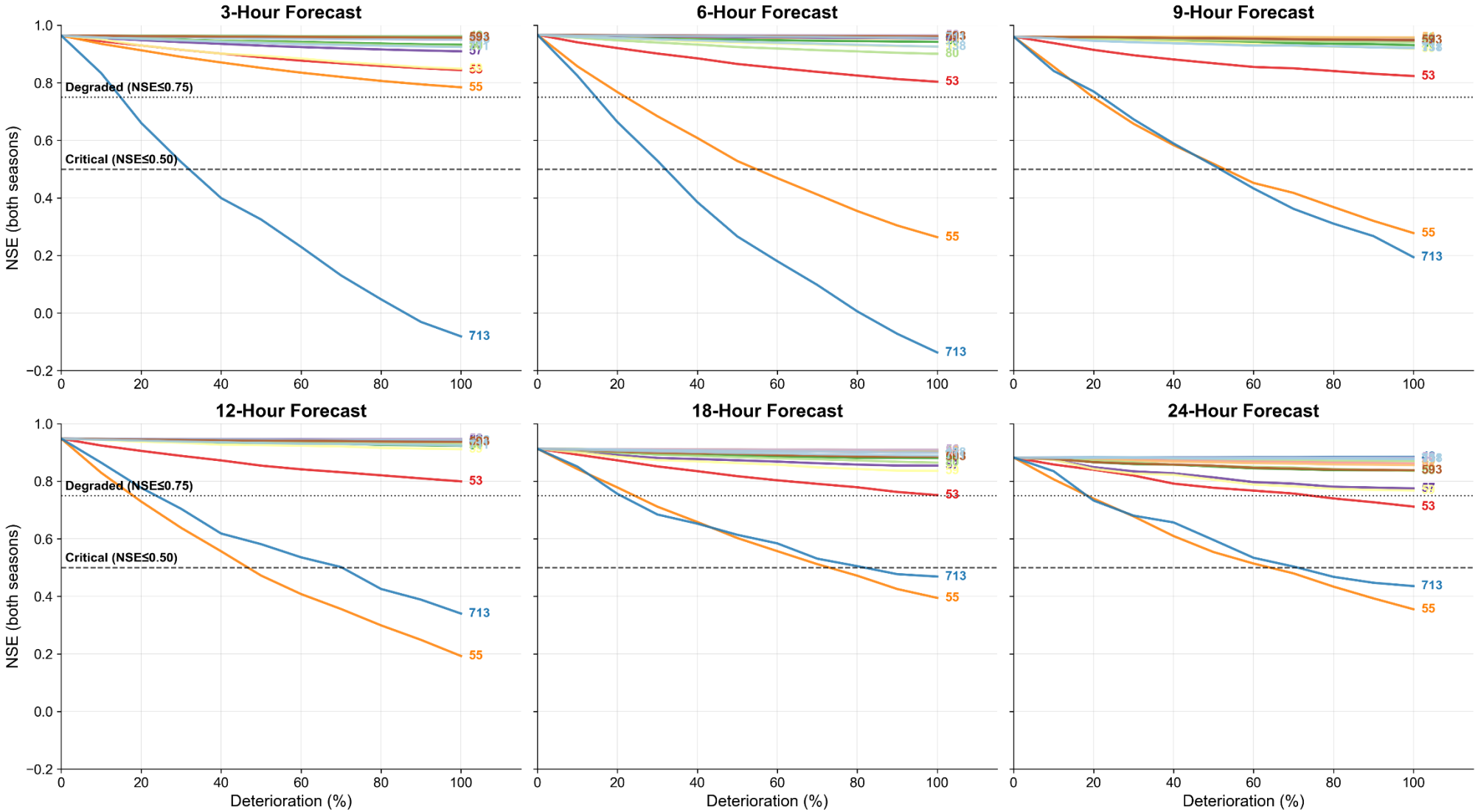
The seasonal deterioration NSE analysis reveals different behaviors in the network's operational resilience for dry and wet periods, driven by the distinct hydrological regimes of the basin. For the wet season (October–March) presented in Figure 35B, the deterioration exposes the fragility of the network. For the 3-hour forecast, the loss of Gauge 713 causes the model to crash to near-zero accuracy (NSE < 0.1). This sensitivity is driven due to the model's missing the hydrograph peaks, characteristic of the season. Without this data, the model loses its hydraulic baseline, compounding errors across successive events. Conversely, during the dry season (April–September)

presented in Figure 36B, the system demonstrates a more robust behavior, with NSE > 0.75 even with high data loss for almost all gauges. However, a notable inversion in sensor hierarchy occurs. Unlike the wet and both seasons, where the proximal Gauge 713 dominates, Gauge 55 emerges as the most critical sensor even at the earliest lead times (3–6 hours).

The KGE deterioration heatmaps for seasonal deterioration also reveals the temporal shift in network dependency, particularly evident at Station 713 (Figure 16). At shorter forecast horizons (3 hours, 6 hours, 9 hours,12 hours), the model exhibits high sensitivity to data availability, with performance degrading rapidly as this feature data is removed. This indicates a low level of redundancy for immediate forecasting; the model is heavily reliant on the closest station's recent auto-regressive data, a feature that constitutes a 'single point of failure' which cannot be easily substituted by other network sensors. Conversely, at longer lead times (18 hours, 24 hours), the model demonstrates significantly higher robustness, maintaining stable KGE scores even under severe data deterioration. This serves as empirical evidence of spatial redundancy within the middle-basin monitoring network. For longer horizons, the hydraulic state of the system is not defined by the local station's immediate history, but by flow propagation dynamics captured by upstream gauges.

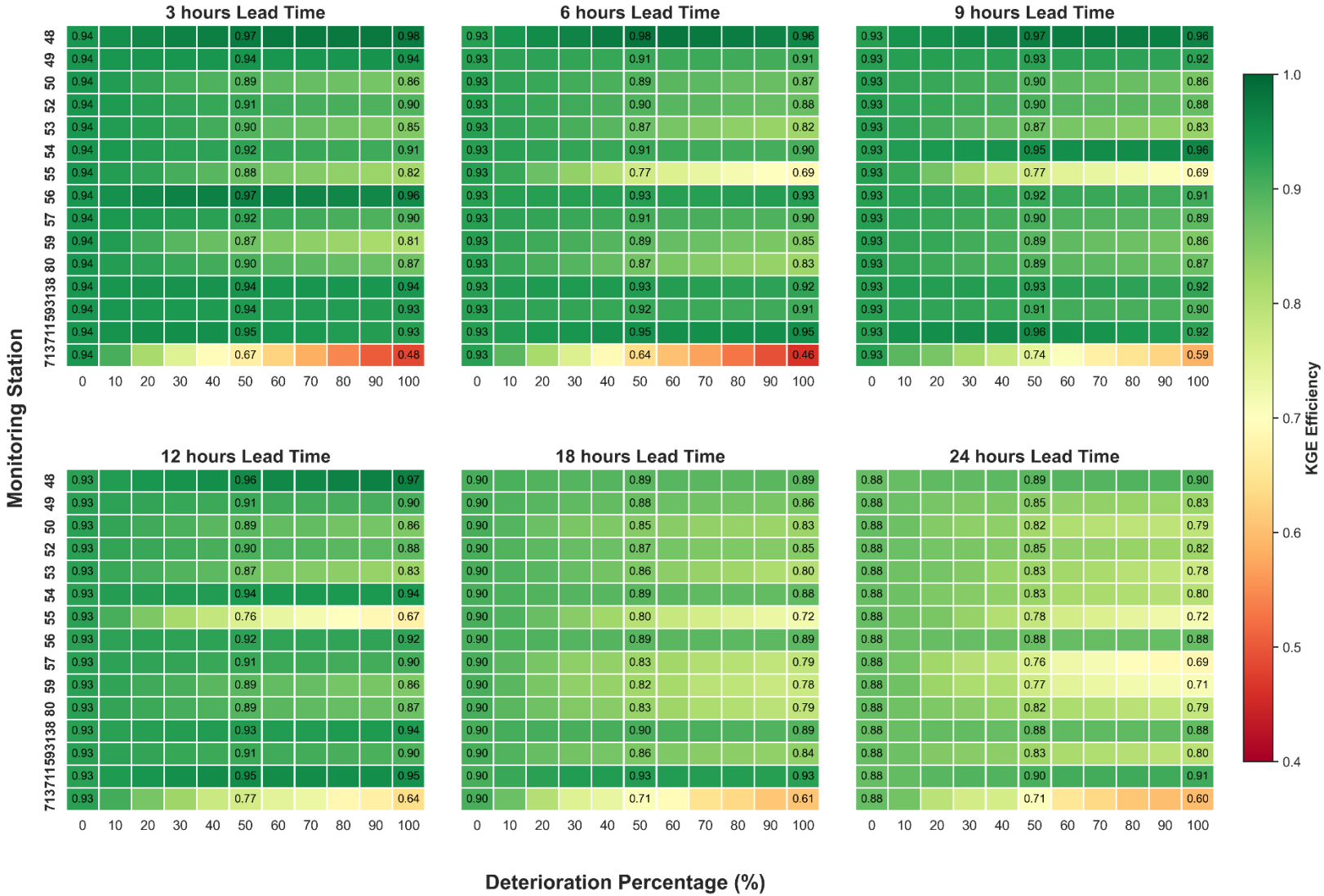
The seasonal analysis regarding the KGE heatmaps reveals that for the dry season, the model doesn't lose accuracy below 0.8 when removing data from gauge 713, even for shorter lead times (3 hours, 6 hours, 9 hours). Instead, the model shifts its attention to gauge 55, in the middle of the basin, losing accuracy to around KGE=0.7. Yet, for the dry season deterioration (Figure 38B) the model still performs well under all deterioration scenarios, since the KGE doesn't drop below 0.7 in any gauge data loss. For the wet season (Figure 37B), the model still relies on gauge 713 for shorter lead times and starts using middle basin data as the lead time increases (12 hours, 18 hours and 24 hours). The metric loss in the wet season deterioration was similar to both seasons' scenario, (KGE >0.4 for 3 hours and KGE >0.6 for 24 hours), but for some gauges data removal the metric slightly increased, indicating that some flood data could be redundant for the model's accuracy or even worsening it at some point.

Figure 15 – NSE Trend across different deterioration scenarios and lead times (Both Seasons).



Source: The author (2025).

Figure 16 – KGE Trend across different deterioration scenarios and lead times (Both Seasons).



Source: The author (2025

To visualize the spatial distribution of these impacts, maps of the basin where the magnitude of the performance drop is represented by the size of the circle at each station's location are presented. Appendix B presents deterioration maps for each lead-time model. The maps show the final NSE value after 10% and 100% data deterioration for the 3-hour (Figure 17 and Figure 18) and 24-hour (Figure 19 and Figure 20) forecasts. The first two maps show that the most significant impact for the 3-hour forecast is concentrated on gauge 713. In contrast, the latter maps illustrate that for the 24-hour forecast, the impact is more distributed across the basin, with a pronounced performance drop associated with the failure of the mid-basin station 55. These maps provide an intuitive geographical summary of which stations are most critical for short-term versus long-term forecasting.

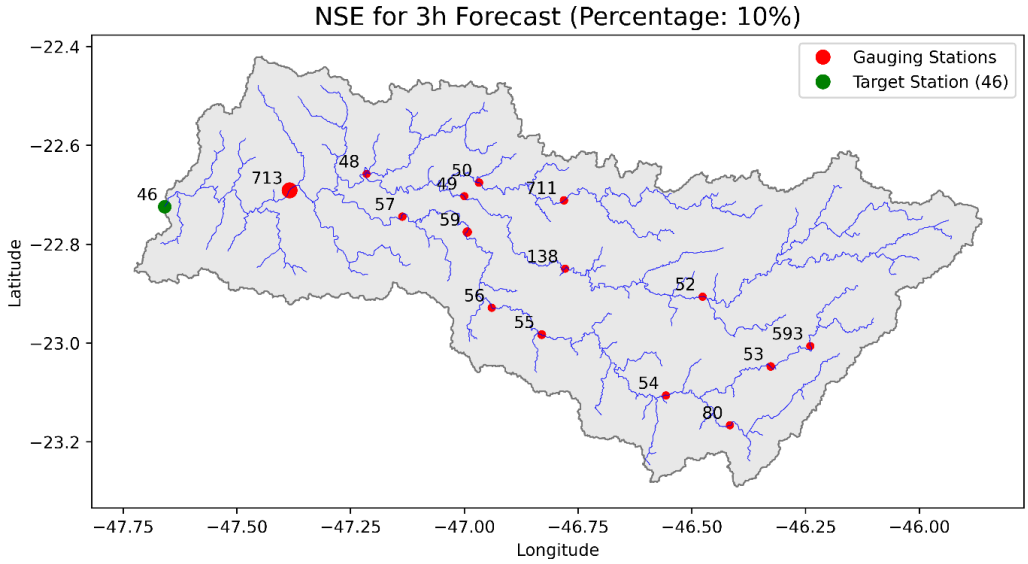


Figure 17 – Spatial representation of model performance (NSE) for 3 hours lead time with 10% data loss.

Source: The author (2025).

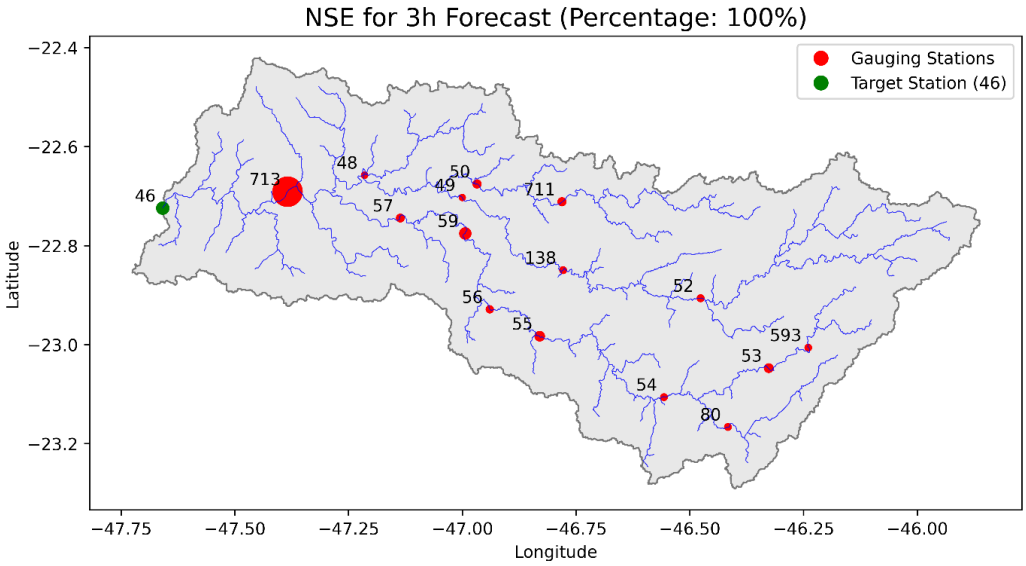


Figure 18 – Spatial representation of model performance (NSE) for 3 hours lead time with 100% data loss.

Source: The author (2025).

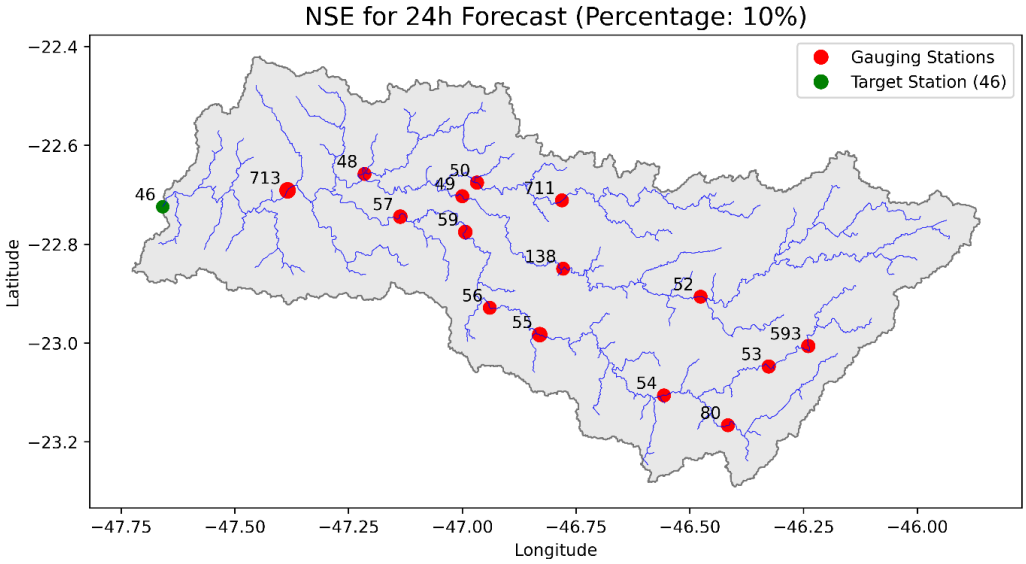


Figure 19 – Spatial representation of model performance (NSE) for 24 hours lead time with 10% data loss.

Source: The author (2025).

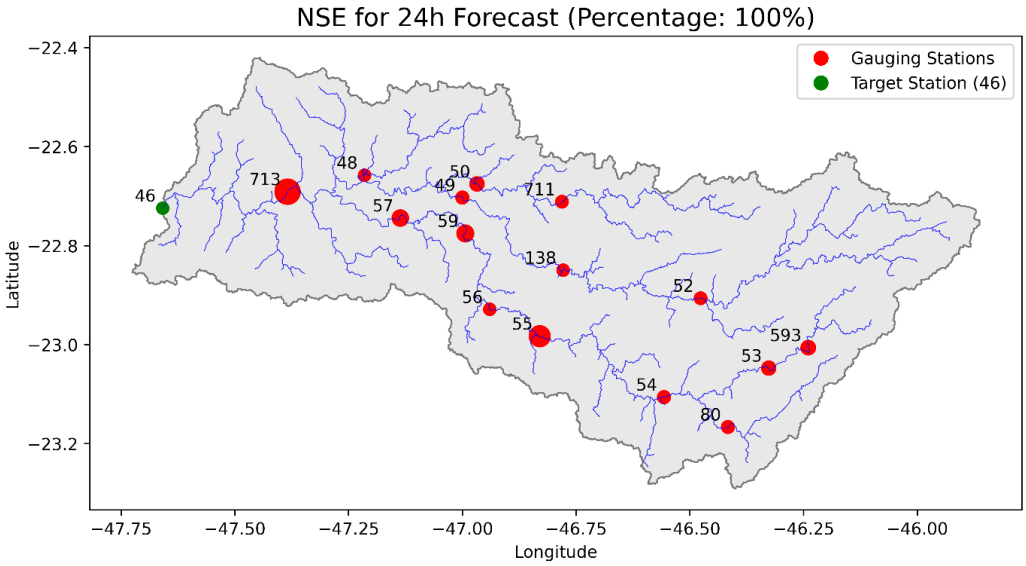


Figure 20 – Spatial representation of model performance (NSE) for 24 hours lead time with 100% data loss.

Source: The author (2025).

5.4 Gauge Importance and Prioritization

To further address model interpretability and validate the results of the deterioration experiments, the SHAP method was employed. This feature attribution analysis provides a direct quantification of each input gauge's contribution to the model's predictions, revealing the internal logic learned by the MLP networks.

It is important to note that the SHAP rankings showed all water-level features to be more important than the rainfall ones, thus corroborating with the assumption made in subchapter 4.4.1. The autocorrelating nature of the water-level dataset made it prevail over the rainfall in terms of feature importance for SHAP evaluation.

The global feature importance for two representative lead times, 3-hour and 24-hour, is presented in the SHAP summary plots in Figure 21 and Figure 22, respectively. These plots rank different input gauges by their SHAP value, representing their overall impact on model output, and visualize the distribution and effect of each gauge's values across all test set predictions. For the 3-hour forecast model, the analysis identifies gauge 713 as the dominant feature. The horizontal spread of its SHAP values is substantially wider than that of any other gauge, indicating its high predictive power. The color mapping reveals a strong positive correlation: high observed water levels at

gauge 713 consistently produce large, positive SHAP values, signifying that these inputs are the primary drivers for high water level predictions at the target station.

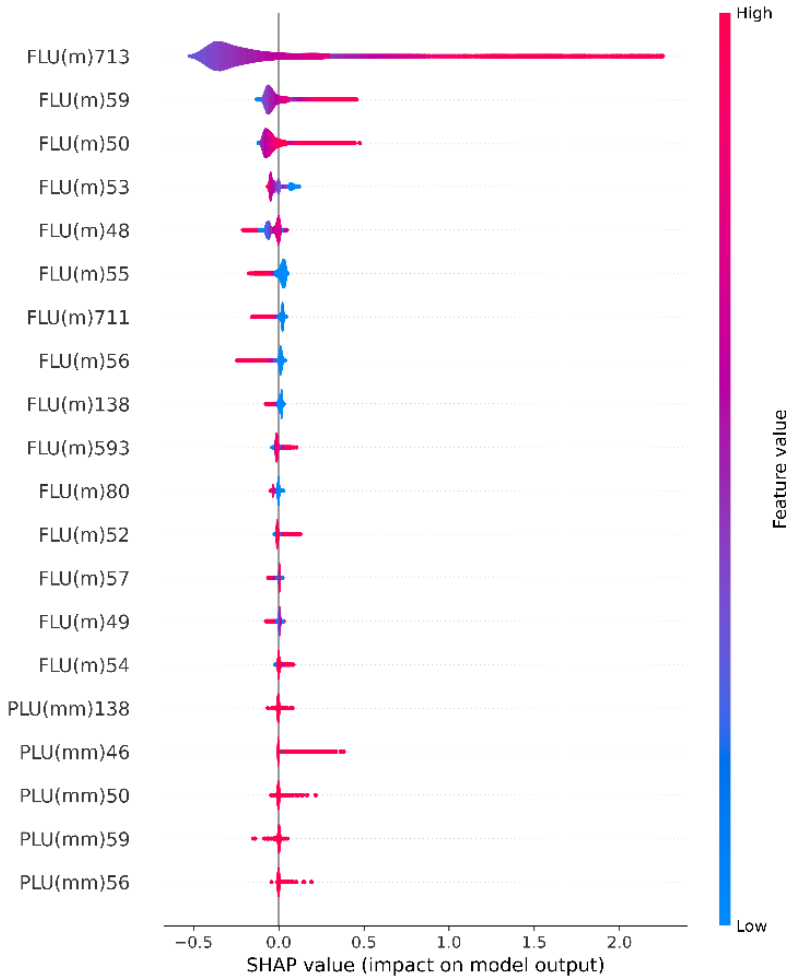


Figure 21 – SHAP values for different input gauges for 3-hour forecast.
Source: The author (2025).

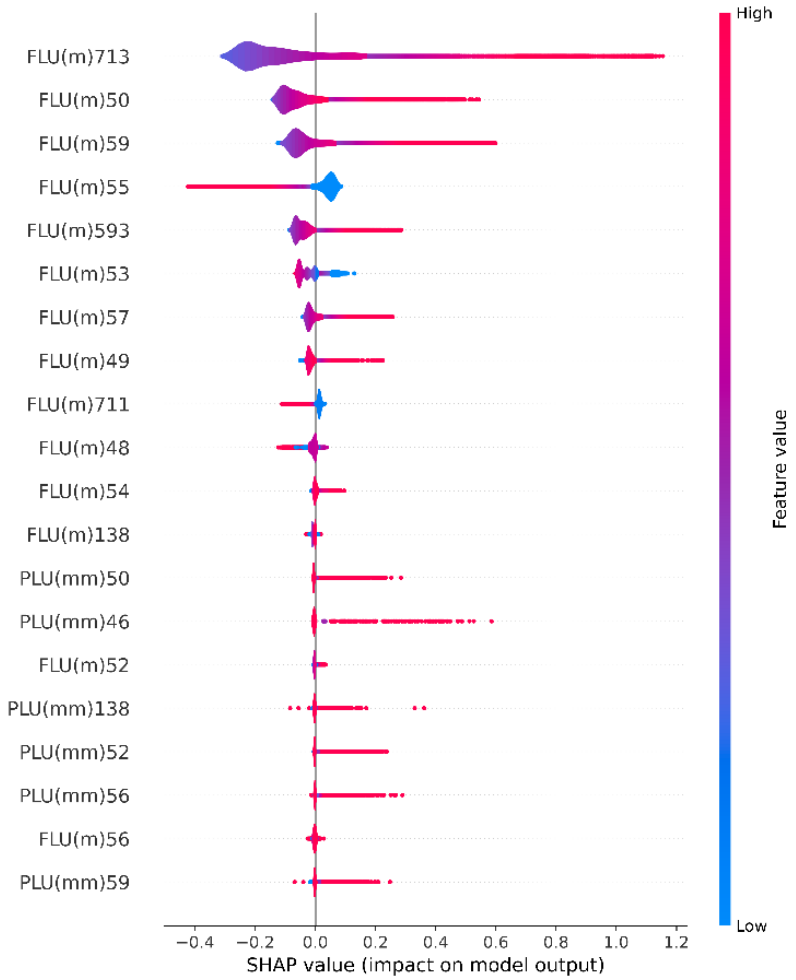


Figure 22 – SHAP values for different input gauges for 24-hour forecast.
Source: The author (2025).

For the 24-hour forecast model, different feature importance is observed. While gauge 713 retains a degree of influence, its ranking diminishes. Instead, mid-basin stations, specifically gauges 55, 59, and 50, emerge as critical features. Gauge 55 is identified as having the highest mean absolute SHAP value for this longer lead time. This result provides an independent line of evidence that corroborates the findings from the deterioration methodology. Additional figures illustrating the results for all lead times are presented in Appendix C. It is possible to note the rising SHAP importance for middle basin gauges (50, 55, 59) as the lead time increases, as expected.

Similar to the data-deterioration maps presented in subchapter 5.3, SHAP value importance maps were plotted. Figure 23 shows the 3-hour lead-time SHAP values per station, while Figure 24 shows the results for 24-hour lead time. Larger circles represent larger SHAP values modules, analogue to the larger circles representing a

broader loss in accuracy in last subchapter maps. The results presented clearly show that for smaller lead times, the closer gauges have greater SHAP values, thus are more important features for flood prediction at the target station. Likewise, the 24-hour map also had a similar result to the ones presented in the deterioration chapter, indicating that both approaches found hydrological reason in the predictions. The maps illustrating all lead times are presented in Appendix D.

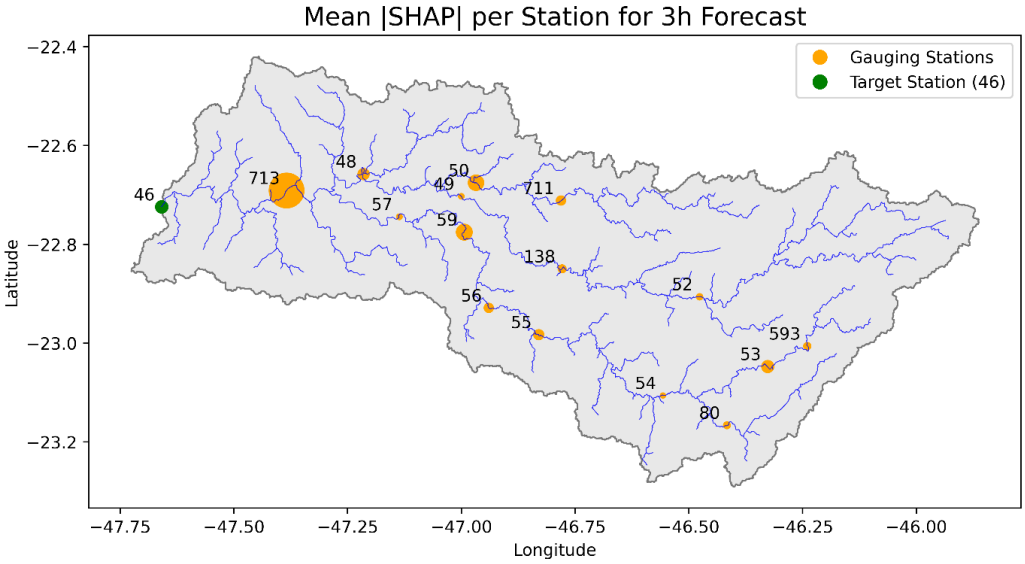


Figure 23 – Spatial representation of SHAP values for 3 hours lead time.
Source: The author (2025).

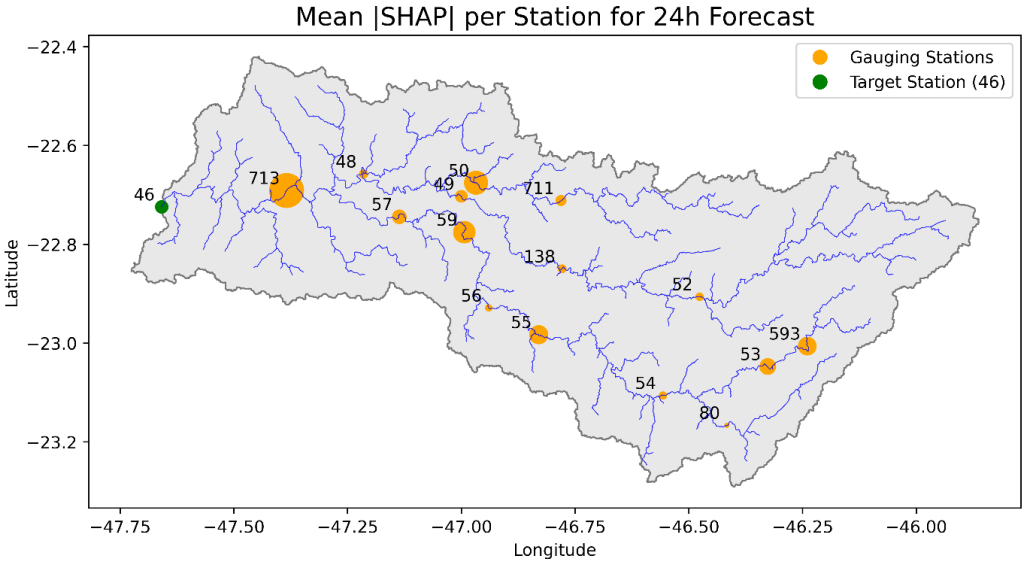


Figure 24 – Spatial representation of SHAP values for 24 hours lead time.
Source: The author (2025).

To verify if the model's feature attribution aligns with physical hydrological processes, the temporal evolution of SHAP values was analyzed alongside the observed hydrograph for the period (January–July 2018), comprising one flood event in January. Figure 25 and Figure 26 illustrate the evolution of SHAP values in the 5 most important stations for each lead time (higher SHAP value in module) and the observed water level data at the target location (Gauge 46). Appendix F presents 6-hour, 9-hour, 12-hour and 18-hour lead-time results.

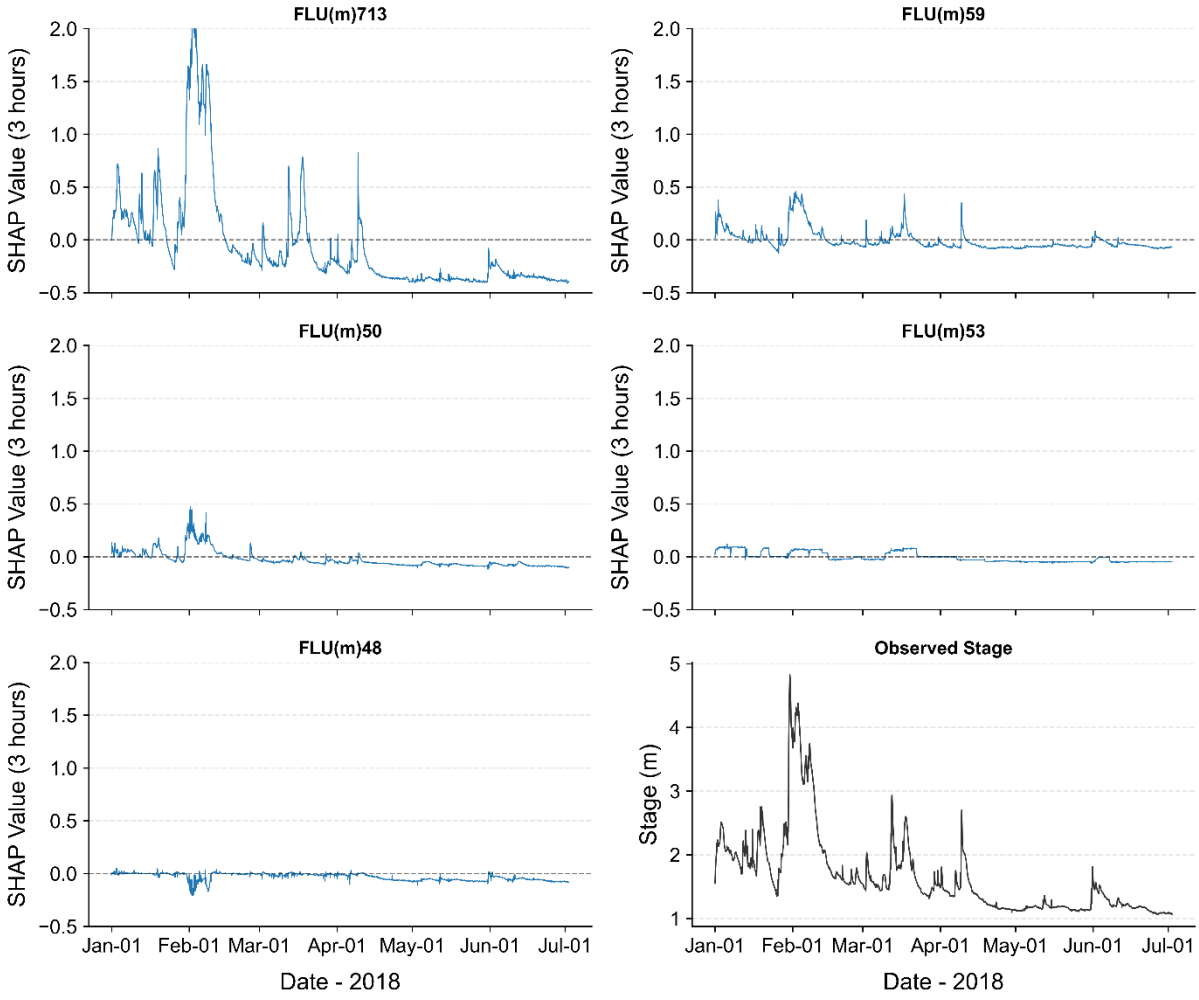


Figure 25 – SHAP Values over time for 3-hour lead time.
Source: The author (2025).

For the 3-hour forecast, the model exhibits a dominant reliance on the nearest upstream sensor, Gauge 713. The SHAP values for this gauge (top left panel) mirror the observed stage hydrograph (bottom right panel) matching almost every peak and recession, reaching contributions exceeding 2.0 during peak flow events. This behavior is hydrologically feasible: given that the flood wave travel time from Gauge 713 is close

to 3 hours, the stage at 713 provides the most immediate and highly correlated signal for a 3-hour prediction. In this horizon, sensors located further upstream (Gauges 48, 50, 53, 59) present smaller SHAP values, smaller than 0.5, indicating they play a secondary role of information rather than primary sources

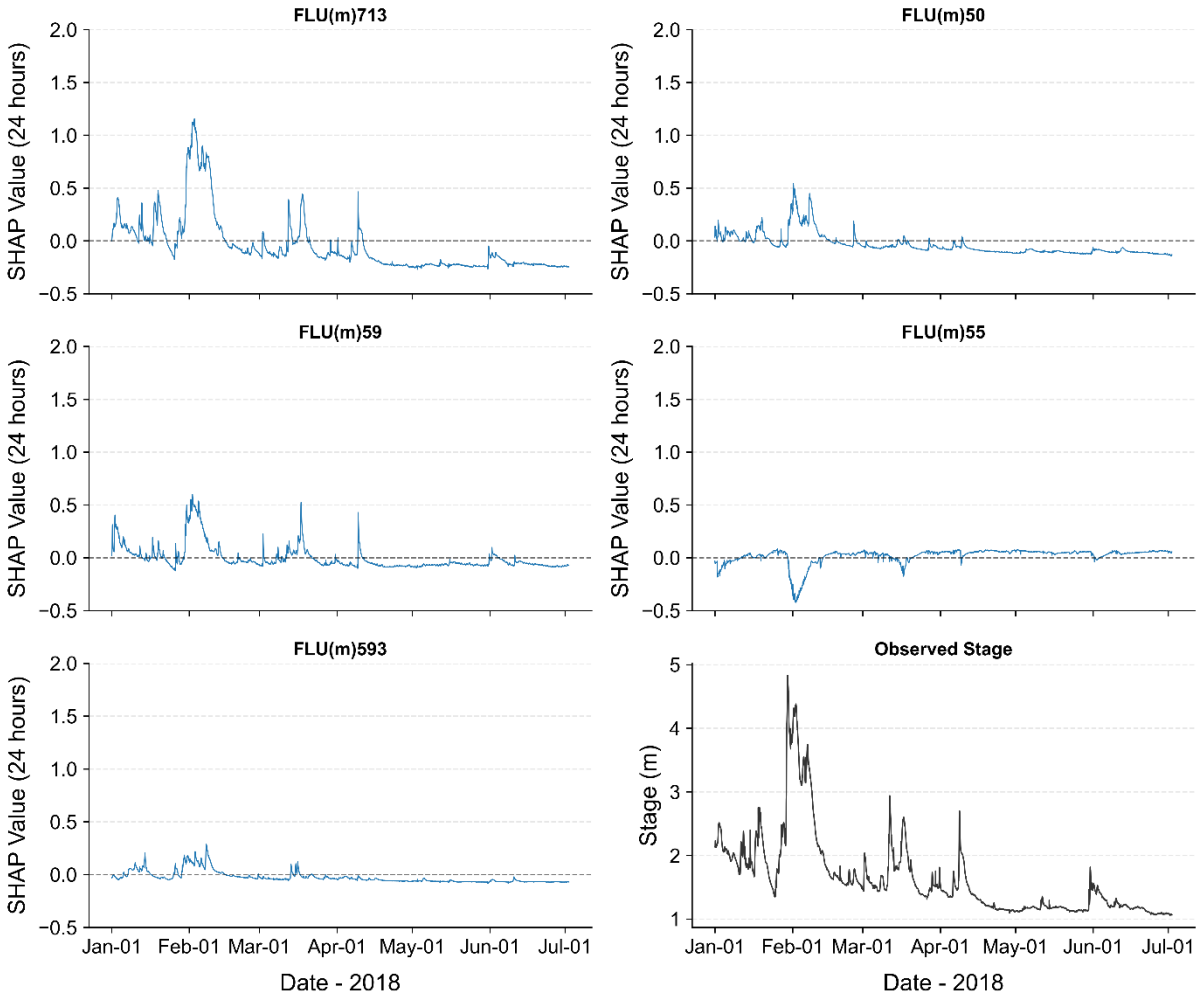


Figure 26 – SHAP Values over time for 24-hour lead time.
Source: The author (2025).

A distinct shift in model logic is observed for the 24-hour forecast. As the prediction horizon extends beyond the immediate travel time of the nearest reach, the model reduces its dependency on Gauge 713, peak SHAP values drop significantly (from >2.0 to ~1.0). Concurrently, the relative importance of further upstream sensors, such as Gauge 50 and Gauge 59, becomes more pronounced. Their SHAP contributions exhibit sharper peaks during flood events compared to the 3-hour model, indicating that the MLP has learned to capture the upstream boundary conditions that will determine the outlet stage 24 hours later.

5.5 Vulnerability Analysis of a Prioritized Network

Aiming to assess the impact of data redundancy on the data-driven framework, the final experimental step evaluates the performance and fragility of the FEWS built only with the most important gauges (gauges 713, 48, 59, 50, and 53). This scenario aims to represent an optimized gauging network, based on the combined results of the deterioration and SHAP analyses, selecting the five most consistently influential gauges. A new MLP model was trained and validated using the data from these five stations as input. This reduced model was then subjected to the same deterioration analysis to quantify its vulnerability.

The baseline performance of this reduced model was slightly lower than the full 15-station model, but it still achieved a high level of accuracy under ideal data conditions. However, the key result of this experiment is the model's heightened sensitivity to data loss. Figure 27 and Figure 28 presents the results of the deterioration experiment on this reduced network for 3-hour lead-time and Figure 29 and Figure 30 for 24-hour lead-time. Both plots show that even a small percentage of data loss (e.g., 10%) from any of the five critical gauges causes a substantially steeper decline in NSE compared to the full model.

For example, when 100% of the data from the most critical station is removed, the model's performance collapses almost entirely, with the NSE dropping to a much lower value than was observed in the full-network experiment under a similar failure. This demonstrates that while a network composed only of the "most important" gauges is efficient and physically sound, it lacks robustness and is less resilient to gauge failure. The absence of redundant or secondary information sources makes the system highly brittle and vulnerable to the failure of even a single critical component. These results provide a quantitative measure of the value of data redundancy in maintaining overall system resilience. The maps presenting results for all lead times are shown in Appendix E.

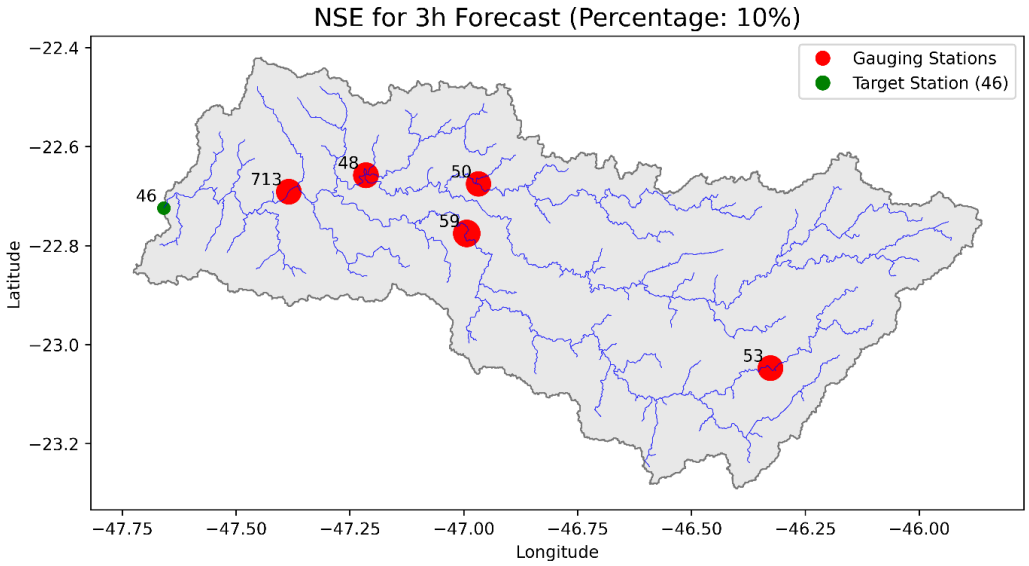


Figure 27 –Spatial representation of model performance (NSE) for 3 hours lead time with 10% data loss.

Source: The author (2025).

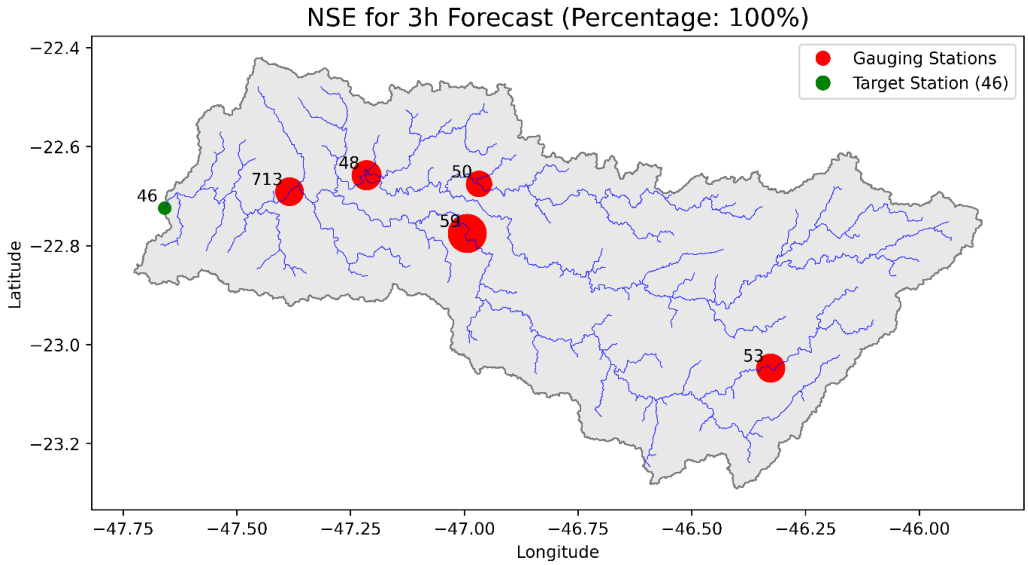


Figure 28 – Spatial representation of model performance (NSE) for 3 hours lead time with 100% data loss.

Source: The author (2025).

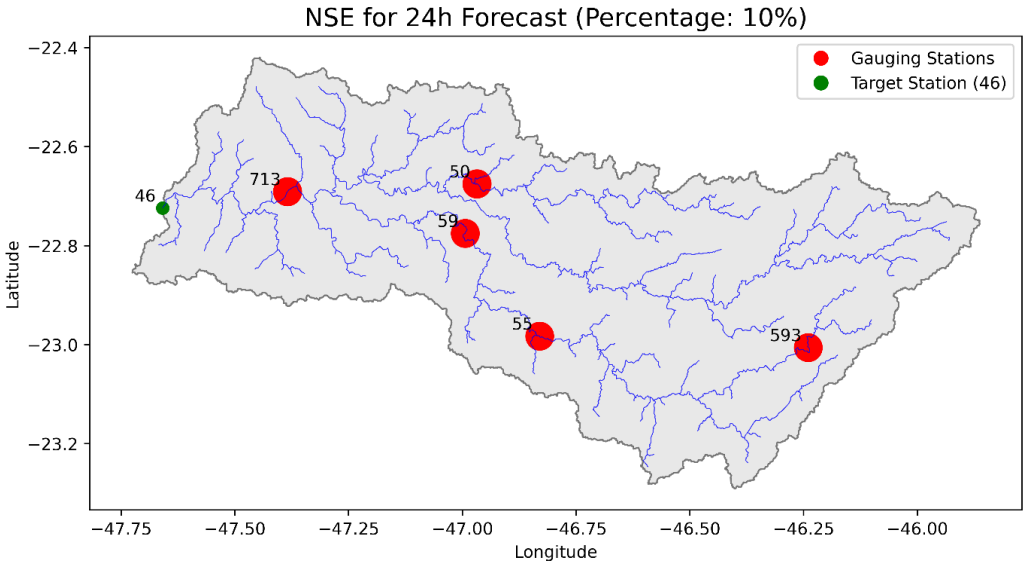


Figure 29 –Spatial representation of model performance (NSE) for 24 hours lead time with 10% data loss.

Source: The author (2025).

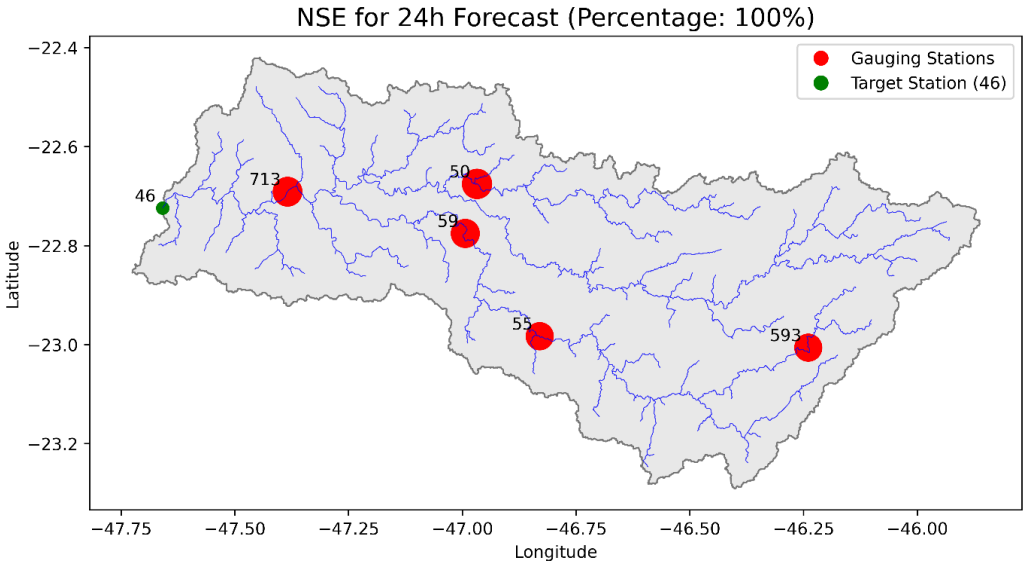


Figure 30 – Spatial representation of model performance (NSE) for 24 hours lead time with 100% data loss.

Source: The author (2025).

6 DISCUSSION

In the present chapter, the findings presented in Chapter 5 are connected to the fundamental hydrological principles and the state-of-the-art literature reviewed in Chapter 3. The objective of this discussion is to explore the broader implications of the results, providing a scientifically grounded explanation for the identified network vulnerabilities and discussing the value of the proposed framework for enhancing the resilience of FEWS.

6.1 Network Resilience Under Stress

The deterioration experiments demonstrated that the flood forecasting system's accuracy is highly sensitive to data failure of individual monitoring gauges. More importantly, the results revealed that the network's vulnerability is not static; the most critical gauges depend on both the forecast lead time and the prevailing seasonal conditions. These findings reflect the model's learning of the underlying hydrological processes that govern flood generation and propagation in the Piracicaba basin.

For short-term forecasts, such as the 3-hour lead time, the analysis consistently identified gauge 713 as the single most critical node in the network. This finding aligns with fundamental principles of flood routing. The 3-hour forecast is dominated by the immediate, local hydrological response and the advection of the flood wave from the closest upstream point. The informational value of more distant gauges is minimal for this short lead time, as the travel time of the flood wave from the upper and middle portions of the basin is significantly longer. The model, therefore, prioritized the most recent, high-relevance information provided by the nearest sensor because the flood event is likely to be at this point 3-hours before reaching the target gauge.

In contrast, for the 24-hour forecast, the system's sensitivity shifted, with gauge 55 in the middle of the basin emerging as a far more influential contributor. This demonstrates that the model has learned a proxy for the physical process of flood wave travel time. To make an accurate 24-hour prediction, the model requires information about the flood wave that is far upstream, allowing sufficient time for that water to concentrate and travel downstream. For these longer lead times, the information chain reaches the mid-basin data to capture the wave's travel time while seemingly capturing

the immediately upstream gauge information to set the magnitude. This behavior, where the window of influence shifts upstream as lead times grow, is consistent with the findings of Menapace et al. (2025) and aligns with the concept that basin response time reflects the propagation of hydraulic potential through the system (Knapp et al., 2025). The model learned that to predict the future, it must look further into the past, both temporally and spatially. This hydrological understanding is reinforced by the SHAP values over time, presented in Appendix F, demonstrating that data-driven models could identify physical contexts in flood forecasting.

Complementing the variance-based assessment provided by NSE, the analysis of the KGE heatmaps offers further insight into the model's degradation. While the NSE metric is highly sensitive to peak mismatches, penalizing the model for timing errors, the KGE decomposition reveals that the system retains significant hydrological consistency even under stress. As observed in the KGE deterioration trends, the decay in performance is notably less precipitous than that of the NSE. This discrepancy indicates that, although the loss of critical upstream inputs (such as Gauge 713 for shorter lead times and 55 for longer lead times) degrades the precise timing of the forecast, the model largely preserves the overall flow variability.

The most meaningful results, however, emerged from the seasonal analysis, which found that the system's performance was more sensitive to data loss during the wet season. This can be explained by the critical role of antecedent moisture conditions and the importance of observing the flow upstream. As shown by Chagas et al. (2022) for southeastern Brazil, flood timing is highly correlated with both rainfall and soil moisture peaks, indicating a system that is primed for a rapid response. During the wet season, high antecedent soil moisture means the basin has a reduced capacity to absorb rainfall, leading to a more direct and less attenuated runoff response. This translates into a quicker hydrograph rise across the basin, speeding up flood routing and improving the forecasting capacity. Consequently, accurate real-time data monitoring during this period is paramount, mainly the spatial distribution of the river flow (Menapace et al., 2025).

Interestingly, the accuracy loss in the dry season is still lower than the case where both seasonal data are missing. The recession periods present in the dry season provide

the model with the necessary antecedent water level for the next storm flood. Gauge 55 emerging as a critical sensor in this season suggests that during low-flow conditions dominated by baseflow recession, the integrated hydrological signal from the mid-basin (Gauge 55) provides a more stable predictor of the system's memory than the potentially erratic signal from the immediate upstream tributary. As noted by Perez et al. (2025), the optimization of reservoir releases in the Piracicaba basin alters the natural flow regime, introducing non-linearities that data-driven models may misinterpret as noise or prediction errors. The watershed response is slower due to the greater infiltration capacity, buffering runoff onset, flow routing, and flood events. This finding underscores that the value of a sensor's data is not constant but is temporally dynamic and conditional on the hydrological state of the basin.

6.2 The Value of Redundancy

While the goal of creating a lean but well-performing network by retaining only the most influential gauges is appealing from a cost-efficiency perspective (Grimaldi et al., 2024), the results demonstrated that such a lean network is also more vulnerable. The model trained on the top five gauges exhibited a considerable drop in performance with even a minor (10%) deterioration in data from one of its gauge inputs. This sensitivity reveals a fundamental trade-off between network efficiency and operational resilience to real world scenarios.

This finding aligns with insights from other hydrological fields, for example, Raleigh et al. (2025) found that while basin-average snowpack information improved water supply forecasts, targeted measurements at specific "snow hotspots" provided consistently greater predictive skill. Their work supports the conclusion that a system's predictive power is not derived from a simple spatial average, but from strategically located points that capture the dominant hydrological signals. Thus, challenging the conventional wisdom of network optimization, suggesting that what might be termed redundancy in a fully functional network could, in fact, be the very source of its resilience. In the full 15-station network, the failure of a single gauge, even a critical one like 713, resulted in a considerable degradation of performance. This is because the model could still draw upon a diverse, albeit incomplete, set of information from the remaining 14 sensors that could have been working as proxy for some hydrological processes. The less important gauges worked as an informational buffer, allowing the system to

continue generating a reasonable forecast even under data failure. In the optimized network, no such buffer existed, thus the failure of one of its five gauges was enough to destabilize the model's prediction.

Therefore, network resilience is a property that arises from the collective information provided by the entire system, not just the sum of its most important parts. The practical implication for water resource managers and agencies responsible for network design is significant (Nguyen et al., 2021), since cutting costs can have a twisted result for forecasting. Decommissioning gauges that appear to have low importance based on an analysis of a complete dataset could inadvertently strip the system of its ability to adapt with the reality of sensor failure and data loss. Instead, a degree of informational redundancy should be viewed as a necessary and valuable investment in the long-term reliability of the FEWS for accurate forecasting.

6.3 Shortcomings and Future Studies

The choice of the MLP as the forecasting model, while sufficient for demonstrating the core concepts of the resilience framework, represents a relatively simple architecture considering the different tools available. The literature review highlighted more advanced models, such as LSTMs and Transformers, which are specifically designed to handle temporal dependencies more explicitly (Tripathy & Mishra, 2024; Frame et al., 2025). It is possible that these more complex architectures could exhibit different sensitivities to data deterioration. Although this study provides a baseline for FEWS and monitoring network resilience, further research is needed to determine if model complexity plays a role in a system's resilience to data loss.

One important improvement to this study lies in the model's handling of rainfall's spatial variability, since the current MLP architecture treats the precipitation time-series from each gauge as an independent input feature. While the model can implicitly learn spatial correlations by analyzing how upstream rainfall and water level patterns lead to a downstream response, it has no explicit knowledge of the spatial rainfall field neither river network connection. As noted in recent literature, GNNs explicitly represent the basin as a graph of connected nodes, an approach that has been shown to improve forecast accuracy by directly modeling the spatial relationships between gauges (Liu et al., 2023). This, along with other architectures like CNNs for handling gridded radar

data (Slater et al., 2025), could be adopted for future refinement of this resilience assessment framework.

The data deterioration algorithm, while systematic, is a representation of real-world sensor failure. Replacing data with zeros simulates a complete loss of signal, which represents a clear "worst-case" scenario, where information is just absent. However, operational sensor failures can manifest in more diverse and complex ways, such as signal noise, intermittent communication dropouts, gradual sensor drift leading to inaccurate readings, or date mismatches along the time-series. The current approach establishes one methodology for testing network response to different percentage of failure, but future studies could build upon this by incorporating different failure modes to better simulate the full spectrum of operational challenges. Some possibilities are applying real problematic data or stochastic simulations.

Furthermore, the temporal extents of the training dataset, while spanning five years of high-resolution data, may not entirely capture the long-term climatic variability in the Piracicaba basin. Hydrological systems are subject to multi-year or decadal cycles of wet and dry periods, often driven by large-scale climate phenomena (Neto et al., 2016). A five-year window, while sufficient for the model to learn event-based rainfall-runoff relationships and seasonal variability, might not be representative of historical extreme events. Consequently, the gauge importance rankings derived from the model could be biased towards the specific hydro-climatic regime of the 2018-2022 period, even though the fitted models are physically sound. The resilience of the network might differ under more extreme drought or flood conditions that were not present in the training data, highlighting the persistent challenge of model generalization under non-stationary climatic conditions (Neto et al., 2016). However, the results observed in the case studies conducted in Italy (Menapace et al., 2025) and Brazil demonstrate the robustness of the methodology under different climatic, hydrological, and topographical conditions. With the extension of the time series, the FEWS resilience should be reassessed to account for a wider range of extreme events.

These limitations do not undermine the central findings of the research but rather frame them within their methodological context, highlighting that the proposed framework provides an extensible methodology that can be adapted and refined in future

research. It’s recommended that future works incorporate greater model complexity and more realistic operational scenarios.

7 CONCLUSION

This thesis successfully developed and validated a novel framework that integrates machine learning, a systematic data deterioration methodology, and XAI to perform a comprehensive resilience assessment of a flood monitoring network. By moving beyond traditional measures of predictive accuracy, this research addressed the critical operational challenge of maintaining forecast reliability in the face of inevitable sensor failures. The application of this framework to the Piracicaba River basin yielded significant insights into the dynamic vulnerabilities of the forecasting system and provided a robust, data-driven method for prioritizing its monitoring gauges.

The experimental results demonstrated that the resilience of the flood monitoring network is not a static property but is indeed dependent on both forecast lead time and seasonal hydrological conditions. The data deterioration experiments revealed that for short-term forecasts (3 hours, 6 hours), the system's accuracy was most vulnerable to the failure of the gauge immediately upstream of the target location. For longer-term forecasts (18 hours, 24 hours), this vulnerability shifted towards the middle of the basin, highlighting that the data-driven model had successfully learned a proxy for the physical process of flood wave propagation. The analysis also showed that the negative impact of sensor failure was more pronounced during the wet season, confirming that the model could have learned the critical importance of antecedent wetness in priming the basin for high-intensity runoff events.

These experimental findings were corroborated by the independent XAI analysis. The application of SHAP revealed that the model's internal processes mirrored closely the results of the deterioration tests, assigning the highest predictive importance to the same gauges identified as critical under the failure scenarios. Finally, the vulnerability analysis of a hypothetical "optimized" network, composed of only the top five most important gauges, provided a clear illustration of the trade-off between efficiency and resilience. While still accurate under ideal conditions, this reduced network proved to be very sensitive to data quality, exhibiting a significance drop in performance from the progressive loss of a single sensor. This finding quantitatively demonstrated that the

informational redundancy provided by the full network is the primary source of its capacity for data degradation resilience.

The research had a general objetive, to evaluate network robustness, that was achieved through the progressive deterioration algorithm. By systematically simulating data loss in increments from 10% to 100% for each gauge, this method provided a direct and quantitative measure of how forecast accuracy, evaluated by the NSE, responds to varying levels of sensor failure. The resulting performance curves served as a clear metric of the system's inability to work under informational loss.

The first specific objective, to analyze the impact of data degradation on the different lead-time models, was addressed by comparing the performance curves of the full 15-station network against the reduced 5-station network. The results clearly demonstrated that the full network could handle degradation for about 50% for the most important gauges, with performance declining gradually as data was removed. In contrast, the “optimized network” proved sensitive to data failure, showing a more proeminent drop in accuracy with the failure of even a single gauge, thus quantifying the value of informational redundancy for a more resilient monitoring network.

The second specific objective, to identify critical gauges, was met through the integrated framework that cross-validated two independent lines of evidence. The empirical impact of sensor failure, from the progressive data deterioration, was correlated with the model's internal reasoning, obtained from the SHAP analysis. The correlation between these two methods, showing that the gauges whose failure caused the largest drop in performance were the mainly the same ones the XAI model assigned the highest feature importance, provided a strong evidence and basis for gauge prioritization.

Finally, the thrid specific objective, to interpret the hydrological significance of the findings, was achieved in the discussion. This was accomplished by connecting the model's learned sensitivities in each lead time to established physical principles and basin, such as the relationship between forecast horizon and flood wave travel time, and the link between seasonal vulnerability and the role of antecedent moisture conditions in runoff generation, ensuring the data-driven results were grounded in physical reality.

This works contributes to the literature is the development and validation of a novel, integrated framework that leverages XAI to move beyond simple gauge prioritization towards a deeper, quantitative assessment of flood monitoring network resilience. While recent studies, such as Menapace et al. (2025) and Raleigh et al. (2025), have established the value of interpretative ML for creating static sensor maintenance rankings, this research provides significant extensions to the state-of-the-art.It shifts the conceptual paradigm from static prioritization to dynamic resilience assessment. By formally defining and measuring properties like robustness and gradual degradation, the framework provides a more holistic understanding of how the forecasting system performs under stress. The study also introduces a more nuanced methodology of progressive, seasonal data deterioration. This moves beyond a simple leave-one-out approach and reveals non-linear performance thresholds, demonstrating that a gauge's importance is not fixed but is conditional on the hydrological state of the basin.By cross-validating the empirical results of these stress tests with the model's internal logic via SHAP, this framework provides a robust, evidence-based method for identifying truly critical infrastructure. Ultimately, this work contributes a more comprehensive and operationally relevant methodology for not only maintaining but also understanding the vulnerabilities of the complex systems that underpin modern flood early warning.

REFERENCES

ADADI, A.; BERRADA, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access, v. 6, p. 52138-52160, 2018.

BAŞAĞAOĞLU, H.; CHAKRABORTY, D.; DO LAGO, C.; GUTIERREZ, L.; ŞAHINLI, M. A.; GIACOMONI, M.; FURL, C.; MIRCHI, A.; MORIASI, D.; ŞENGÖR, S. S. A. Review on Interpretable and Explainable Artificial Intelligence in Hydroclimatic Applications. Water, v. 14, n. 8, p. 1230, 2022.

BAYATI, A.; AMELI, A. A.; RAZAVI, S. Evaluating the functional realism of deep learning rainfall–runoff models using catchment hydrology principles. Water Resources Research, v. 62, n. 1, e2025WR040076, 2025.

BHASME, P.; VAGADIYA, J.; BHATIA, U. Enhancing predictive skills in physically-consistent way: Physics Informed Machine Learning for hydrological processes. Journal of Hydrology, v. 615, p. 128618, 2022.

BINETTI, M. S.; CAMPANALE, C.; MASSARELLI, C.; URICCHIO, V. F. The Use of Weather Radar Data: Possibilities, Challenges and Advanced Applications. Earth, v. 3, n. 1, p. 157-171, 2022.

BISHOP, C. M. Neural networks for pattern recognition. Oxford: Oxford University Press, 1995.

BRANISAVLJEVIĆ, N.; PRODANOVIĆ, D.; ARSIĆ, M.; SIMIĆ, Z.; BOROTA, J. Hydro-Meteorological Data Quality Assurance and Improvement. Journal of the Serbian Society for Computational Mechanics, v. 3, n. 1, p. 228-249, 2009.

CEA, L.; COSTABILE, P. Flood Risk in Urban Areas: Modelling, Management and Adaptation to Climate Change. A Review. Hydrology, v. 9, n. 3, p. 50, 2022.

CEA, L.; GARRIDO, M.; PUERTAS, J. Experimental validation of two-dimensional depth-averaged models for forecasting rainfall–runoff from precipitation data in urban areas. Journal of Hydrology, v. 382, n. 1-4, p. 88-102, 2010.

CHAGAS, V. B. P.; CHAFFE, P. L. B.; BLÖSCHL, G. (2022). Process controls on flood seasonality in Brazil. Geophysical Research Letters, 49, e2021GL096754.

CHEN, J.-A.; NIU, W.; REN, B.; WANG, Y.; SHEN, X. Survey: Exploiting Data Redundancy for Optimization of Deep Learning. ACM Computing Surveys, New York, v. 55, n. 10, p. 1–38, 2023.

COLLINS, M. L.; KAPUCU, N. Early warning systems and disaster preparedness and response in local government. Disaster Prevention and Management, v. 17, n. 5, p. 587-600, 2008.

DE LUCA, D. L.; ALLAMANO, P.; BORGA, M.; DEIDDA, R.; MUGNAI, A.; NAKAJO, T.; PUCA, S.; RATTO, C.; RIVOLTA, G.; ROTH, G.; SICCARDI, F. Rainfall nowcasting models: state of the art and possible future perspectives. Hydrological Sciences Journal, v. 70, n. 9, p. 1419-1438, 2025.

DOTTORI, F.; SZEWCZYK, W.; CISCAR, J. C.; ZHAO, F.; ALFIERI, L.; HIRABAYASHI, Y.; BIANCHI, A.; MONGELLI, I.; FRIELER, K.; BETTS, R. A. and FEYEN, L. Increased human and economic losses from river flooding with anthropogenic warming. Nature Climate Change, [S.l.], v. 8, p. 781–786, 2018.

FERRAZ, F. F.; FERRAZ, E. S. B.; BALLESTER, M. V. R.; MORAES, J. M.; VICTORIA, R. L.; MARTINELLI, L. A. Previsão de áreas inundadas na cidade de Piracicaba (SP) através de Sistema de Informações Geográficas (SIG). Revista Brasileira de Recursos Hídricos, v. 3, n. 3, p. 17-27, 1998.

FRAME, J. M.; KRATZERT, F.; KLOTZ, D.; GAUCH, M.; SHALEV, G.; GILON, O.; QUALLS, L. M.; GUPTA, H. V.; NEARING, G. S. Deep learning rainfall-runoff predictions of extreme events. Hydrology and Earth System Sciences, v. 26, n. 12, p. 3377–3392, 2022. DOI: 10.5194/hess-26-3377-2022.

FRAME, J. M.; ARAKI, R.; BHUIYAN, S. A.; BINDAS, T.; RAPP, J.; BOLOTIN, L.; DEARDORFF, E.; LIU, Q.; HACES-GARCIA, F.; LIAO, M.; FRAZIER, N.; and OGDEN, F. L.: Machine Learning for a Heterogeneous Water Modeling Framework, JAWRA Journal of the American Water Resources Association, 61, e70 000, 2025.

GAUCH, M.; KLOTZ, D.; KRATZERT, F.; NEARING, G.; HOCHREITER, S.; LIN, J. A Machine Learner's Guide to Streamflow Prediction. In: AI for Earth Sciences Workshop at NeurIPS, 2020.

GAUCH, M.; KRATZERT, F.; KLOTZ, D.; NEARING, G.; COHEN, D.; GILON, O. How to deal w___ missing input data. Preprint, 2025.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. Deep learning. Cambridge: MIT Press, 2016.

GLOROT, X., BENGIO, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS 2010).

GRIMALDI, S. et al. Optimizing sensor location for the parsimonious design of flood early warning systems. Journal of Hydrology X, v. 24, p. 100182, 2024.

GUNNING, D.; STEFIK, M.; CHOI, J.; MILLER, T.; STUMPF, S.; YANG, G.-Z. XAI—Explainable artificial intelligence. Science Robotics, v. 4, n. 37, eaay7120, 2019.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. The elements of statistical learning: data mining, inference, and prediction. 2. ed. New York: Springer, 2009.

HAN, D.; KWONG, T.; LI, S. Uncertainties in real-time flood forecasting with neural networks. Hydrological Processes, v. 21, p. 223–228, 2007.

HAYKIN, S. Neural networks and learning machines. 3. ed. Upper Saddle River: Pearson, 2009.

HAYDER, I. M.; AL-AMIEDY, T. A.; GHABAN, W.; SAEED, F.; NASSER, M.; AL-ALI, G. A.; YOUNIS, H. A. An Intelligent Early Flood Forecasting and Prediction Leveraging Machine and Deep Learning Algorithms with Advanced Alert System. *Processes*, Basel, v. 11, n. 2, p. 481, Feb. 2023.

HORNIK, K.; STINCHCOMBE, M.; WHITE, H. Multilayer feedforward networks are universal approximators. Neural Networks, v. 2, n. 5, p. 359-366, 1989.

HOSSEINI, F.; PRIETO, C.; ÁLVAREZ, C. An explainable AI approach for interpreting regionally optimized deep neural networks in hydrological prediction. Journal of Hydrology, v. 661, 133689, 2025. DOI: 10.1016/j.jhydrol.2025.133689.

INTERGOVERNMENTAL PANEL ON CLIMATE CHANGE (IPCC). Climate Change 2021: The Physical Science Basis: Contribution of Working Group I to the Sixth Assessment Report of the IPCC. Cambridge: Cambridge University Press, 2021.

JUNG, D.; AXELSSON, D. A Study on Redundancy and Intrinsic Dimension for Data-Driven Fault Diagnosis. In: INTERNATIONAL CONFERENCE ON PRINCIPLES OF DIAGNOSIS AND RESILIENT SYSTEMS (DX), 35., 2024. Proceedings... Dagstuhl: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2024. v. 125, p. 4:1–4:17. (OASIcs).

KINGMA, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. arXiv preprint arXiv:1412.6980.

KINGSTON, G. B.; MAIER, H. R.; LAMBERT, M. F. Calibration and validation of neural networks to ensure physically plausible hydrological modeling. Journal of Hydrology, v. 314, n. 1-4, p. 158-176, 2005.

KIRCHNER, J. W. Getting the right answers for the right reasons: linking measurements, analyses, and models to advance the science of hydrology. Water Resources Research, v. 42, W03S04, 2006.

KNAPP, J. L. A., BERGHUIJS, W. R., FLORIANCIC, M. G., & KIRCHNER, J. W. (2025). Catchment hydrological response and transport are affected differently by precipitation intensity and antecedent wetness. Hydrology and Earth System Sciences, 29, 3673–3685.

KORNELSEN, K.; COULIBALY, P. Comparison of interpolation, statistical, and data-driven methods for imputation of missing values in a distributed soil moisture dataset. *Journal of Hydrologic Engineering*, v. 19, n. 1, p. 26-43, 2014.

KRATZERT, F., KLOTZ, D., SHALEV, G., KLAMBAUER, G., HOCHREITER, S., and NEARING, G.: Towards learning universal, regional, and local hydrological behaviors

via machine learning applied to large-sample datasets, Hydrology and Earth System Sciences, 23, 5089–5110, 2019.

KUHANESWARAN, B.; SORWAR, G.; ALAEI, A. R.; TONG, F. Evolution of Data-Driven Flood Forecasting: Trends, Technologies, and Gaps—A Systematic Mapping Study. Water, v. 17, n. 15, p. 2281, 2025.

KUNDZEWICZ, Z. W. et al. Flood Risk and Climate Change: Global and Regional Perspectives. *Hydrological Sciences Journal*, v. 59, n. 1, p. 1–28, Jan. 2013.

LI, K.; PERSAUD, D.; CHOUDHARY, K.; DECOST, B.; GREENWOOD, M.; HATTRICK-SIMPERS, J. Exploiting redundancy in large materials datasets for efficient machine learning with less data. Nature Communications, London, v. 14, n. 7283, p. 1-10, 2023.

LIU, G.; OUYANG, S.; QIN, H.; LIU, S.; SHEN, Q.; QU, Y.; ZHENG, Z.; SUN, H.; ZHOU, J. Assessing spatial connectivity effects on daily streamflow forecasting using Bayesian-based graph neural network. Science of the Total Environment, v. 855, p. 158968, 2023.

LUNDBERG, S. M.; LEE, S. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, v. 30, p. 4765–4774, 2017.

NASH, J. E.; SUTCLIFFE, J. V. River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology*, v. 10, n. 3, p. 282–290, 1970.

NEARING, G.; COHEN, D.; DUBE, V.; GAUCH, M.; GILON, O.; HARRIGAN, S.; HASSIDIM, A.; KLOTZ, D.; KRATZERT, F.; METZGER, A.; NEVO, S.; PAPPENBERGER, F.; PRUDHOMME, C.; SHALEV, G.; SHENZIS, S.; TEKALIGN, T.; WEITZNER, D.; MATIAS, Y. AI Increases Global Access to Reliable Flood Forecasts. arXiv preprint arXiv:2307.16104, 2023.

NETO, A. R.; DA PAZ, A. R.; MARENGO, J. A.; CHOU, S. C. Hydrological processes and climate change in hydrographic regions of Brazil. Journal of Water Resource and Protection, v. 8, n. 12, p. 1103–1127, 2016.

NGUYEN, L. D.; NGUYEN, H. T.; DANG, P. D. N.; DUONG, T. Q.; NGUYEN, L. K. Design of an automatic hydro-meteorological observation network for a real-time flood warning system: a case study of Vu Gia-Thu Bon river basin, Vietnam. Journal of Hydroinformatics, v. 23, n. 2, p. 324-334, 2021.

MACHADO, R.E., LOPES, T.R., DUARTE, S.N., 2025. Projected climate and land-use change impacts on streamflow: The case study of Piracicaba basin – Brazil. International Journal of River Basin Management, 1–16.

MAIER, H. R.; TAGHIKHAH, F. R.; NABAVI, E.; RAZAVI, S.; GUPTA, H.; WU, W.; RADFORD, D. A. G.; HUANG, J. How much X is in XAI: Responsible use of “Explainable” artificial intelligence in hydrology and water resources. Journal of Hydrology X, v. 25, 100185, 2024.

MENAPACE, A.; RODRIGUES, A. F.; DALLA TORRE, D.; LARCHER, M.; HERRERA, M.; BRENTAN, B. Sensors prioritisation for hydrological forecasting based on interpretable machine learning. Journal of Hydrology, v. 663, 2025.

MILLER, J. D.; HUTCHINS, M. The impacts of urbanization and climate change on urban flooding and urban water quality: A review of the evidence concerning the United Kingdom. Journal of Hydrology: Regional Studies, v. 12, p. 345–362, ago. 2017.

MIMIKOU, M. A.; BALTAS, E. A. Flash flood forecasting using weather radar. In: Hydraulic Engineering Software V. Southampton: WIT Press, 1994. p. 297-304.

MONTÁNS, F. J.; CHINESTA, F.; GÓMEZ-BOMBARELLI, R.; KUTZ, J. N. Data-driven modeling and learning in science and engineering. Comptes Rendus Mécanique, Paris, v. 347, n. 11, p. 845–855, 2019.

PARKER, D. J.; TUNSTALL, S. M.; MCCARTHY, S. M. New insights into the benefits of flood warnings: Results from a household survey in England and Wales. Environmental Hazards, v. 7, p. 193-210, 2007.

PEREZ, R. F.; TERCINI, J. R. B.; HOSSODA, D. H.; RABIOGLIO, V. L. G.; BONNECARRÈRE, J. I. Advancing Water Resources Management Through Reservoir

Release Optimization: A Study Case in Piracicaba River Basin in Brazil. Hydrology, v. 12, n. 10, 269, 2025.

PLATE, E. J. Flood risk and flood management. Journal of Hydrology, [S.l.], v. 267, p. 2–11, 2002. DOI: 10.1016/S0022-1694(02)00135-X.

RALEIGH, M. S.; SMALL, E. E.; BAIR, E. H.; WOBUS, C.; RITTGER, K. Snow monitoring at strategic locations improves water supply forecasting more than basin-wide mapping. Communications Earth & Environment, v. 6, n. 665, 2025.

RASHEED, Z.; ARAVAMUDAN, A.; ZHANG, X.; ANAGNOSTOPOULOS, G. C.; NIKOLOPOULOS, E. I. Combining global precipitation data and machine learning to predict flood peaks in ungauged areas with similar climate. Advances in Water Resources, v. 192, 2024. Art. 104781.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016. p. 1135-1144.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. Nature, v. 323, n. 6088, p. 533–536, 1986.

SAMEK, W.; MONTAVON, G.; VEDALDI, A.; HANSEN, K. H.; MÜLLER, K. R. (Eds.). Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Cham: Springer, 2019.

SCHALKOFF, R. J. (1997). Artificial Neural Networks. McGraw-Hill.

SCHANZE, J. Flood risk management – A basic framework. In: SCHANZE, J.; ZEMAN, E.; MARSCHAL, J. (org.). Flood risk management: Hazards, vulnerability and mitigation measures. Dordrecht: Springer, 2006. p. 1–20.

SLATER, L.; BLOUGOURAS, G.; DENG, L.; DENG, Q.; FORD, E.; HOEK VAN DIJKE, A.; HUANG, F.; JIANG, S.; LIU, Y.; MOULDS, S.; SCHEPEN, A.; YIN, J.; ZHANG, B.

Challenges and opportunities of ML and explainable AI in large-sample hydrology. Philosophical Transactions of the Royal Society A, v. 383, 20240287, 2025.

STAUDINGER, M.; HERZOG, A.; LORITZ, Ralf; HOUSKA, T.; POOL, S.; SPIELER, D.; WAGNER, P. D.; MAI, J.; KIESEL, J.; THOBER, S.; GUSE, B.; EHRET, U. How well do process-based and data-driven hydrological models learn from limited discharge data? Hydrology and Earth System Sciences, v. 29, p. 5005–5029, 2025.

SUNKPHO, J.; OOTAMAKORN, C. Real-time flood monitoring and warning system. Songklanakarin Journal of Science and Technology, v. 33, n. 2, p. 227-235, 2011.

THAKKER, D.; MISHRA, B. K.; ABDULLATIF, A.; MAZUMDAR, S.; SIMPSON, S. Explainable Artificial Intelligence for Developing Smart Cities Solutions. Smart Cities, v. 3, 2020.

TRIPATHY, K. P.; MISHRA, A. K. Deep learning in hydrology and water resources disciplines: concepts, methods, applications, and research directions. Journal of Hydrology, v. 628, p. 130458, 2024.

UNITED NATIONS INTERNATIONAL STRATEGY FOR DISASTER REDUCTION (UNISDR). UNISDR Terminology on Disaster Risk Reduction. Geneva, 2009.

XU, Y.; LIN, K.; HU, C.; WANG, S.; WU, Q.; ZHANG, L.; RAN, G. Deep transfer learning based on transformer for flood forecasting in data-sparse basins. Journal of Hydrology, v. 625, 2023.

YANTO, M.; LIVNEH, B.; RAJAGOPALAN, B.; KASPRZYK, J. Hydrological model application under data scarcity for multiple watersheds, Java Island, Indonesia. Journal of Hydrology: Regional Studies, v. 9, p. 127-139, 2017.

YUAN, X.; WOOD, E. F.; MA, Z. A review on climate-model-based seasonal hydrologic forecasting: physical understanding and system development. WIREs Water, v. 2, n. 5, p. 523-536, 2015.

ZHAO, Y.; WU, X.; ZHANG, W.; LAN, P.; QIN, G.; LI, X.; LI, H. A deep learning-based probabilistic approach to flash flood warnings in mountainous catchments. Journal of Hydrology, v. 652, 132677, 2025.

ZHANG, J.; CAO, C.; NAN, T.; JU, L.; ZHOU, H.; ZENG, L. A Novel Deep Learning Approach for Data Assimilation of Complex Hydrological Systems. Water Resources Research, v. 60, 2024.

APPENDIX A – Hydrometeorological Time Series

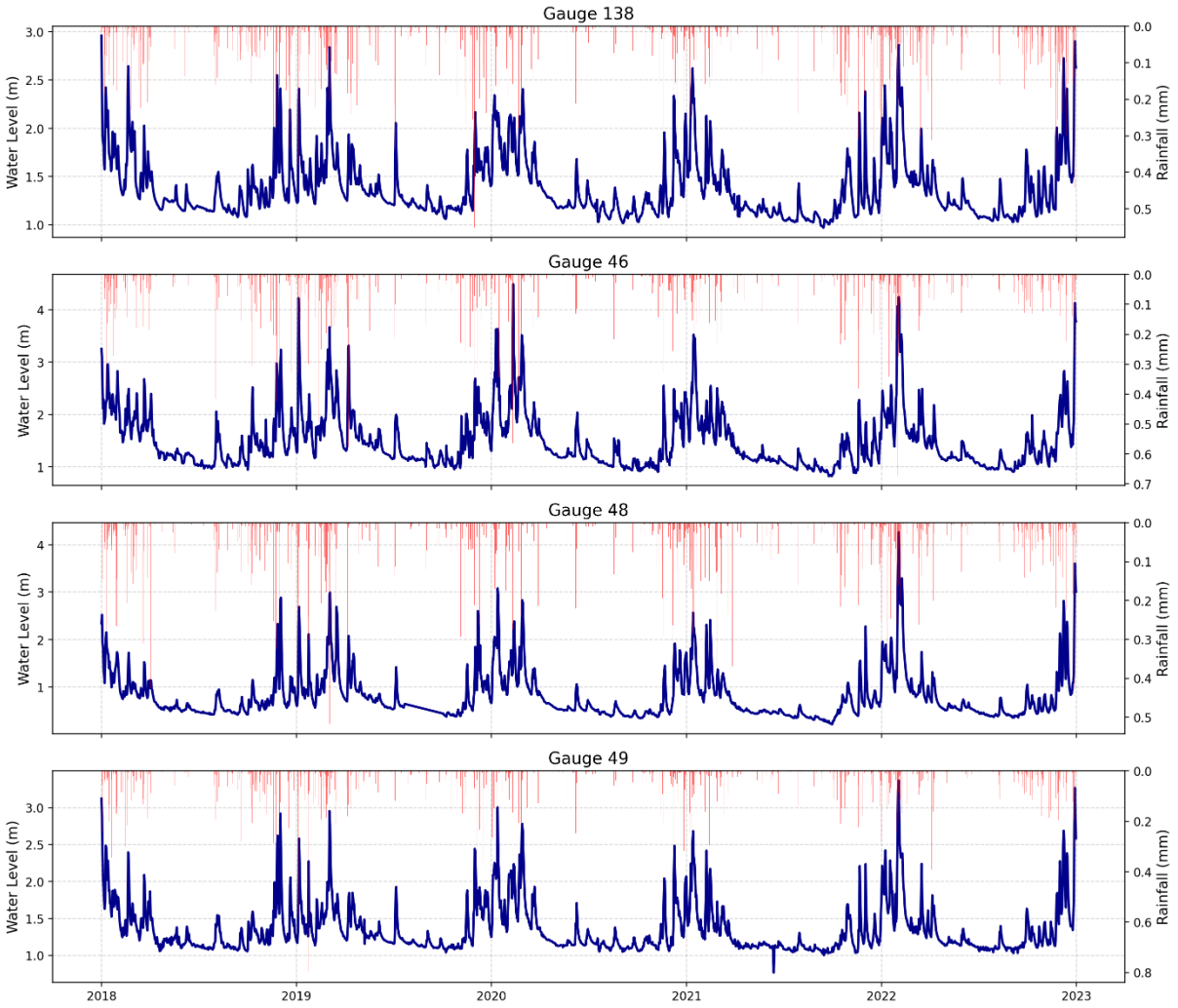


Figure 31A – Water level and rainfall data for gauges 138, 46, 48 and 49

Source: The author (2025).

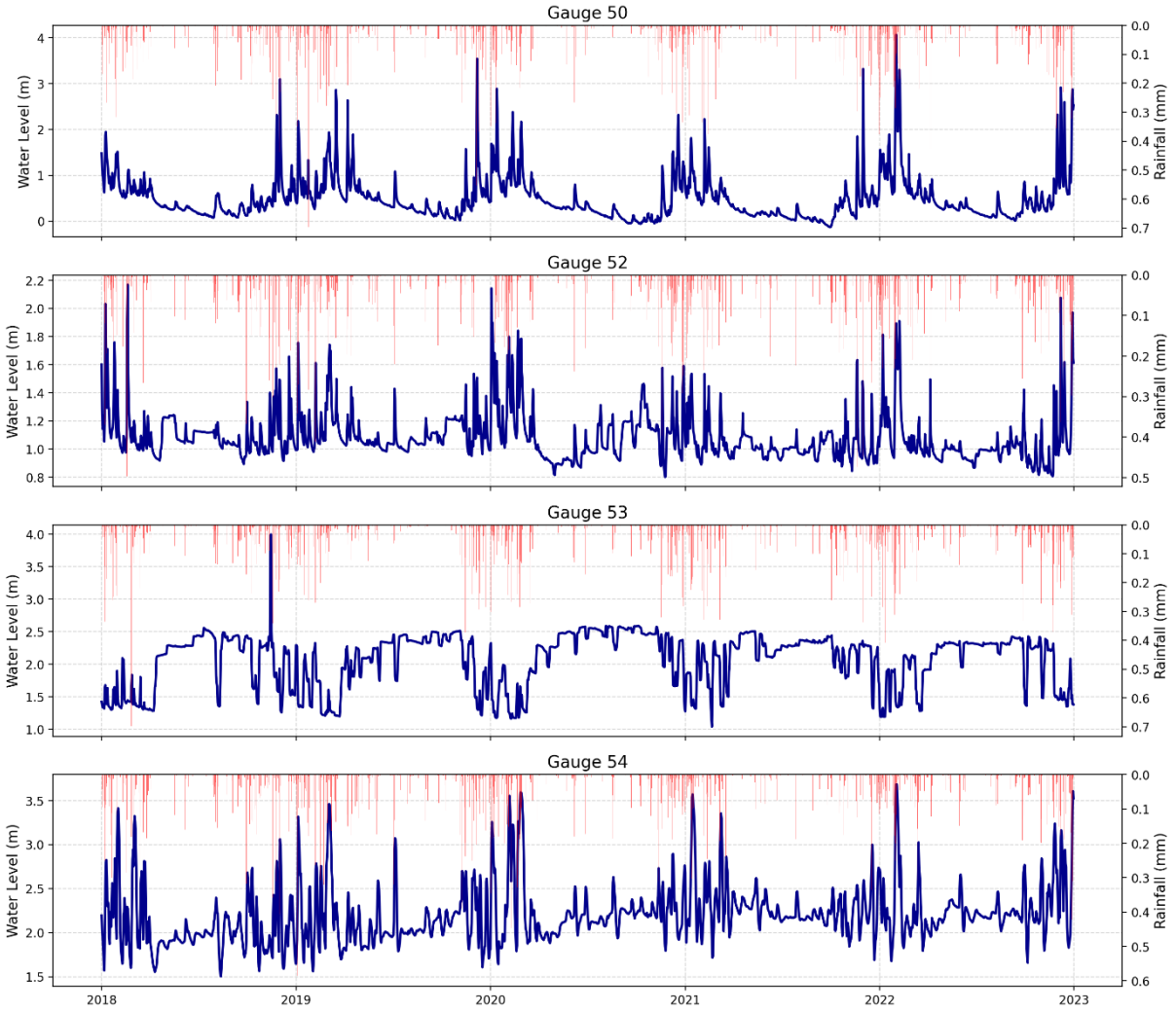


Figure 32A – Water level and rainfall data for gauges 50, 52, 53 and 54

Source: The author (2025).

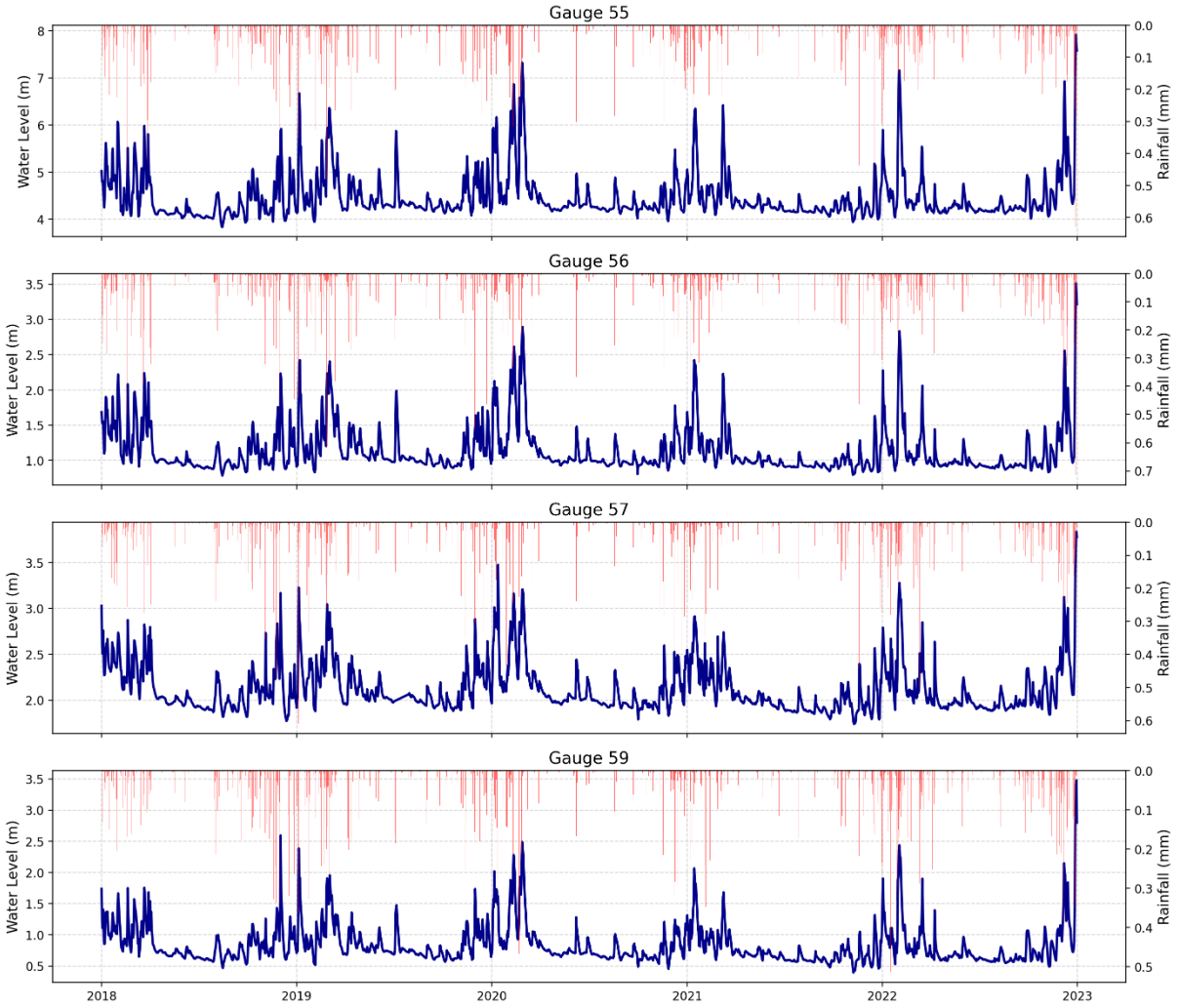


Figure 33A – Water level and rainfall data for gauges 55, 56, 57 and 59

Source: The author (2025).

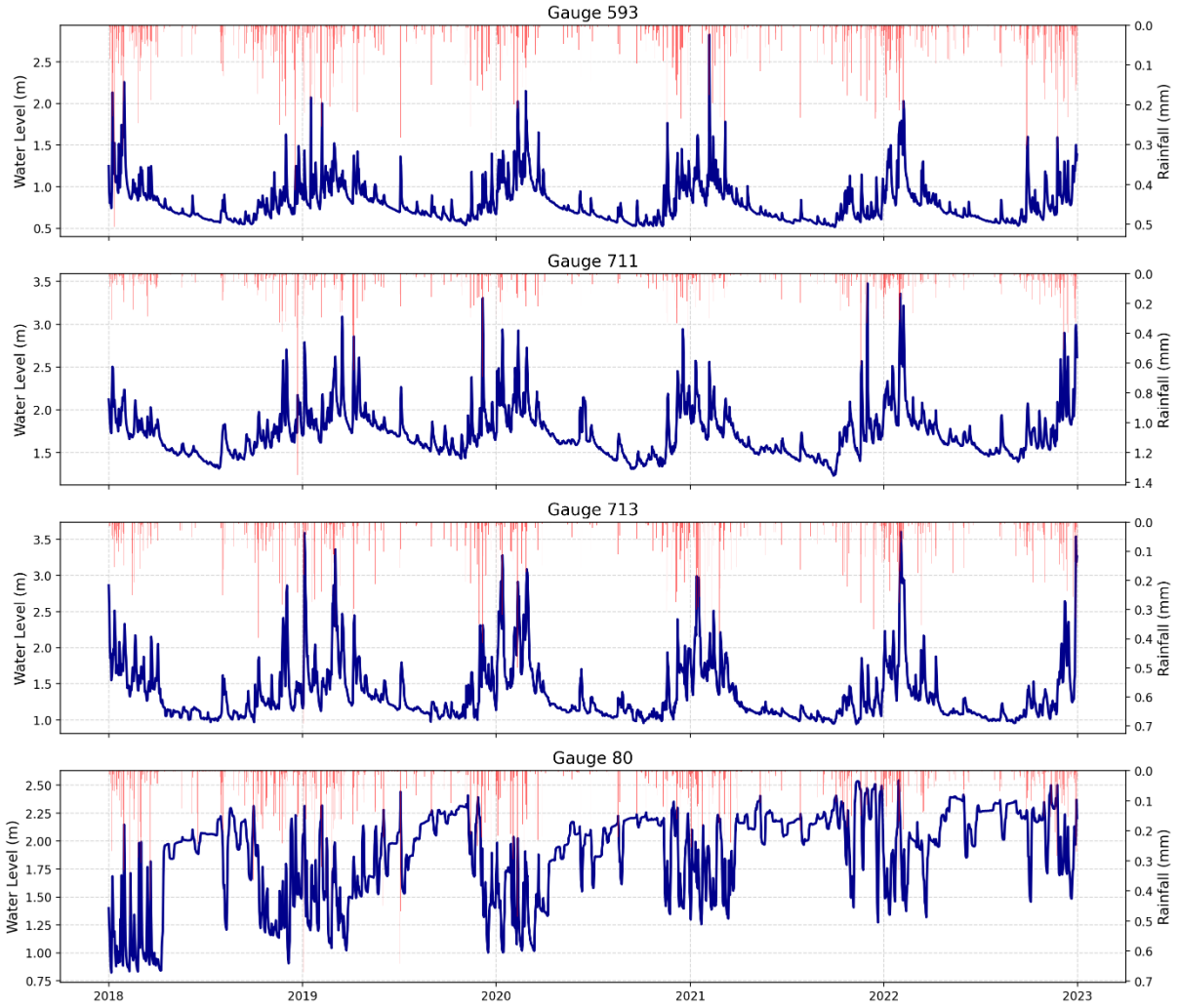


Figure 34A – Water level and rainfall data for gauges 593, 711, 713 and 80

Source: The author (2025).

APPENDIX B – Deteriorated Metric Trends (Seasonal)

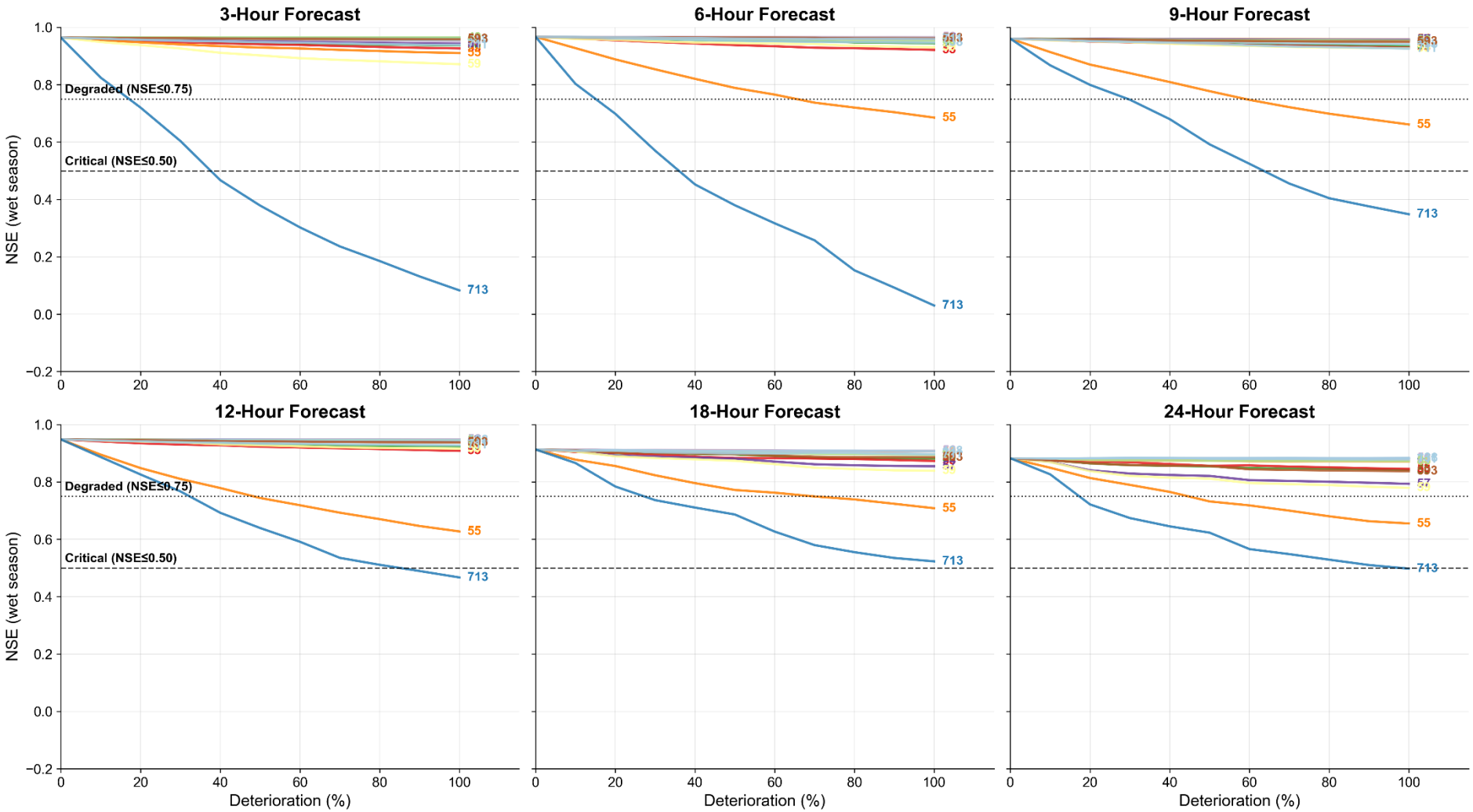


Figure 35B – NSE Trend across different deterioration scenarios and lead times (Wet Season).

Source: The author (2025).

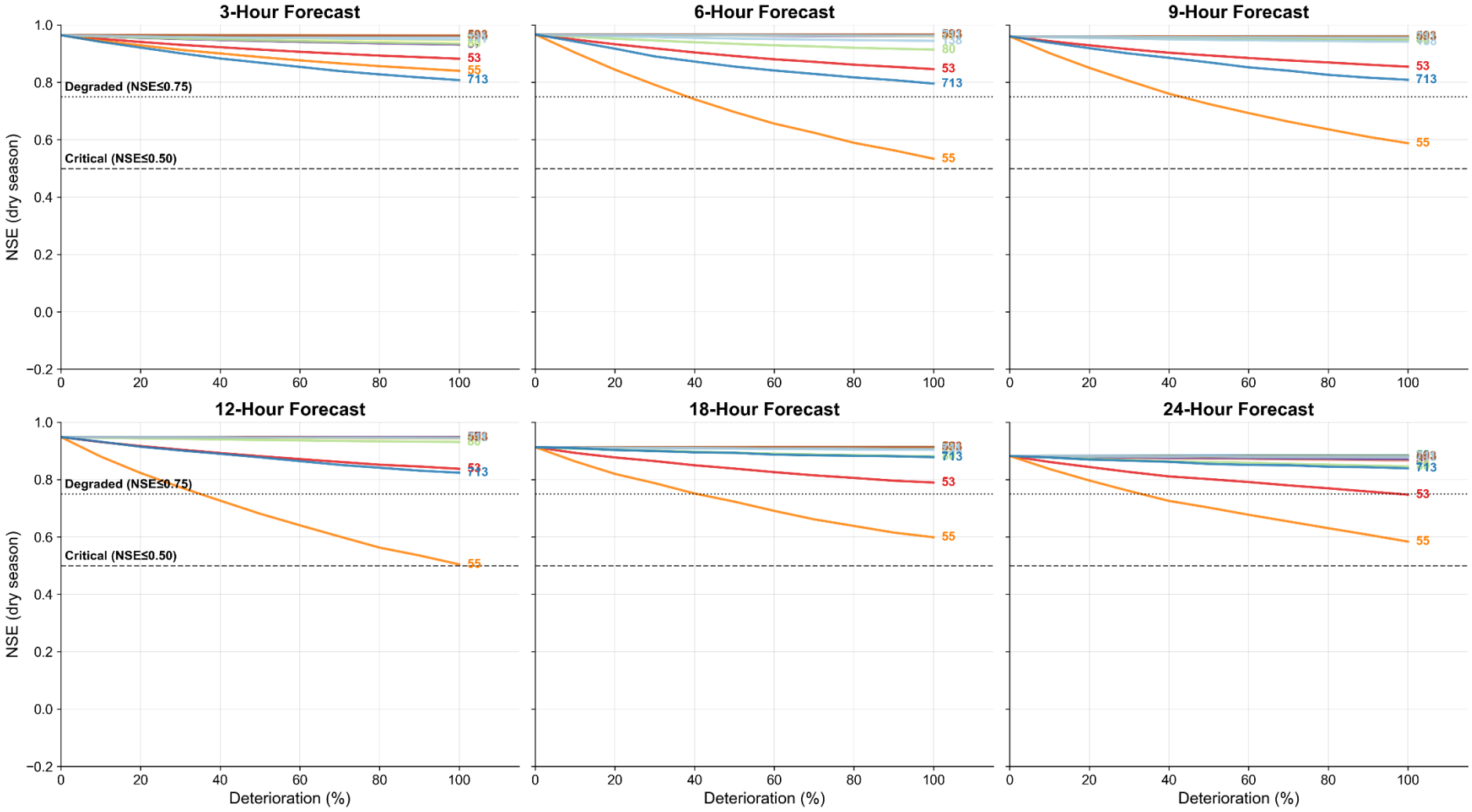


Figure 36B – NSE Trend across different deterioration scenarios and lead times (Dry Season).

Source: The author (2025).

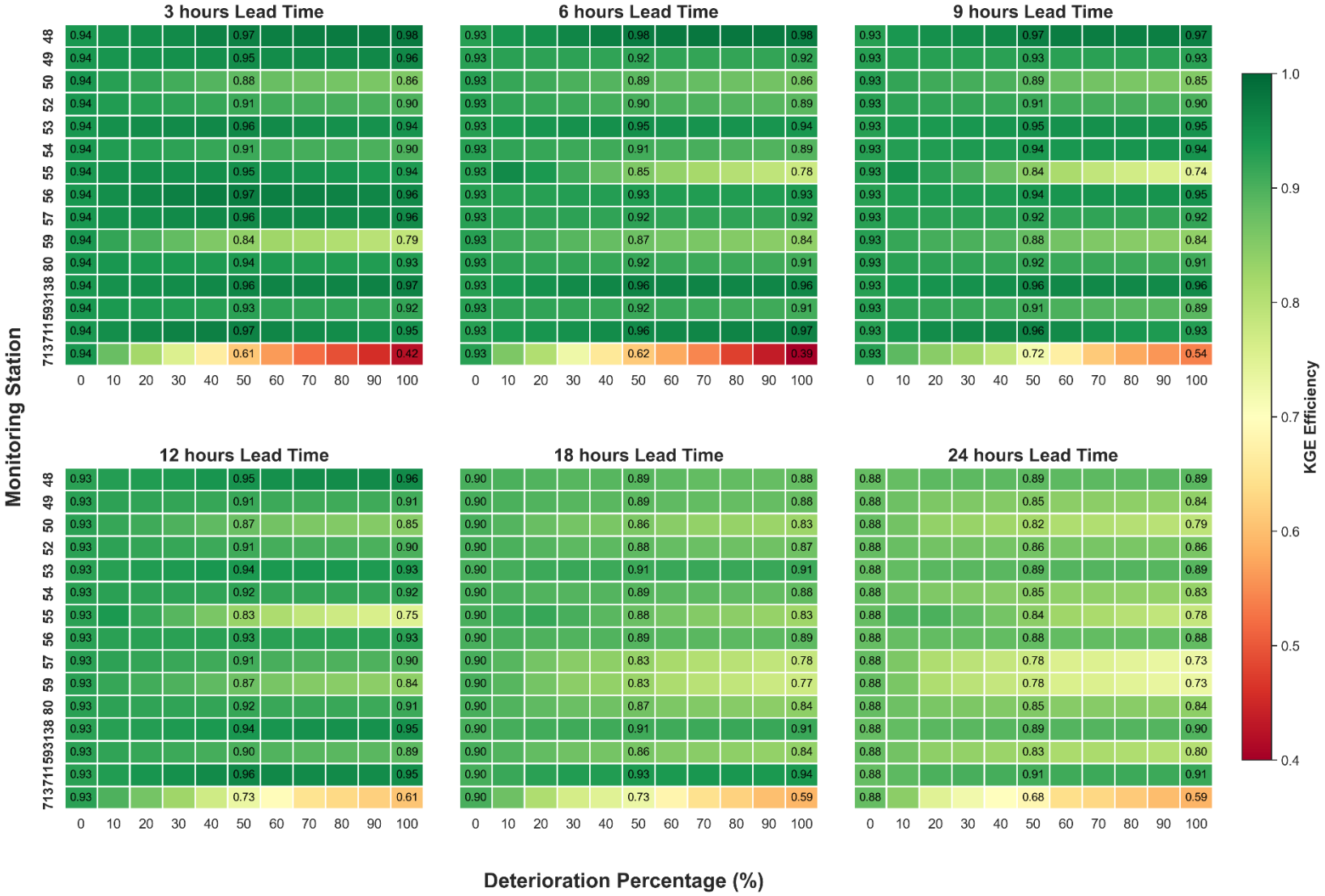


Figure 37B – KGE Trend across different deterioration scenarios and lead times (Wet Season).

Source: The author (2025).

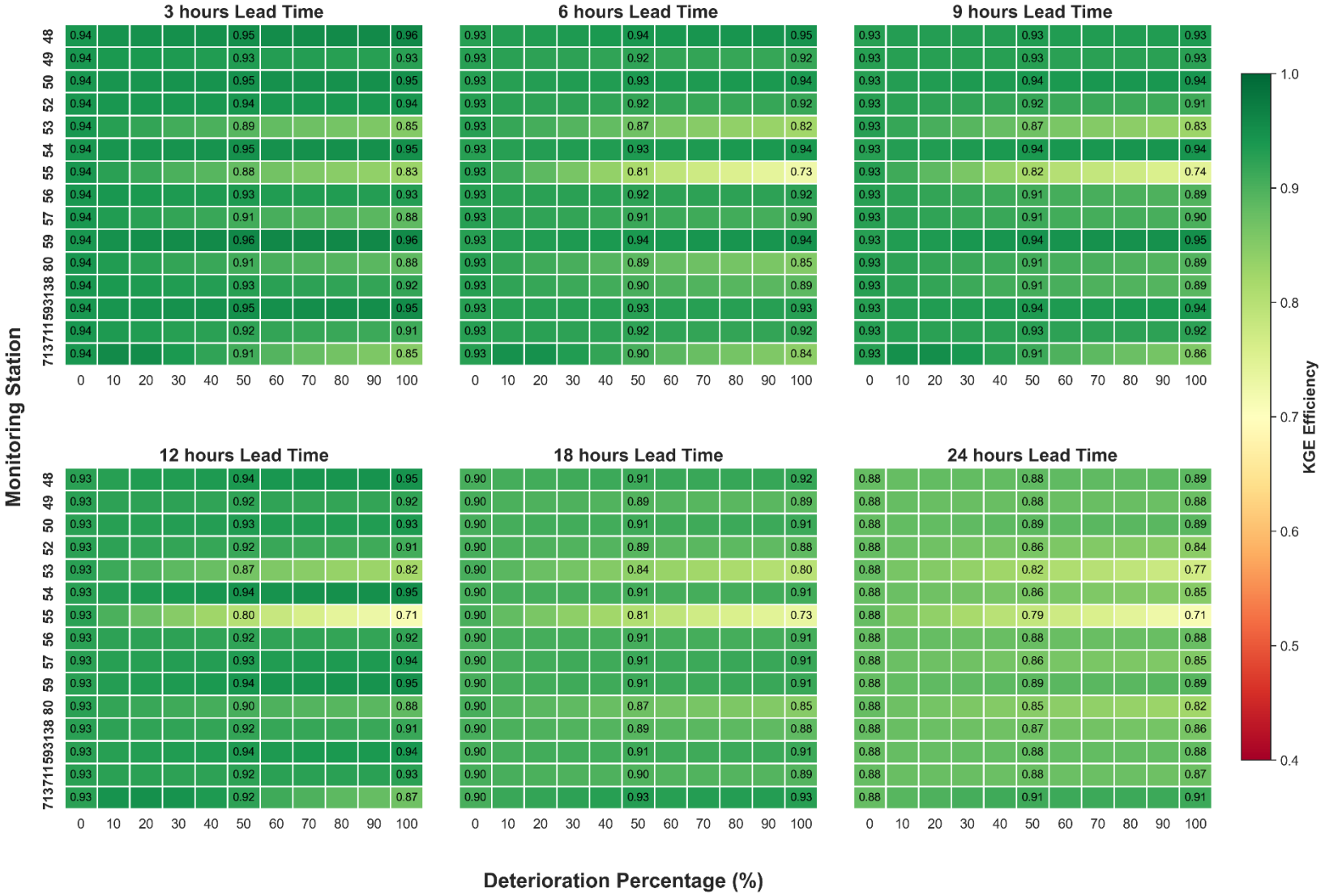


Figure 38B – KGE Trend across different deterioration scenarios and lead times (Dry Season).

Source: The author (2025).

APPENDIX C – Network Data-Deterioration Maps

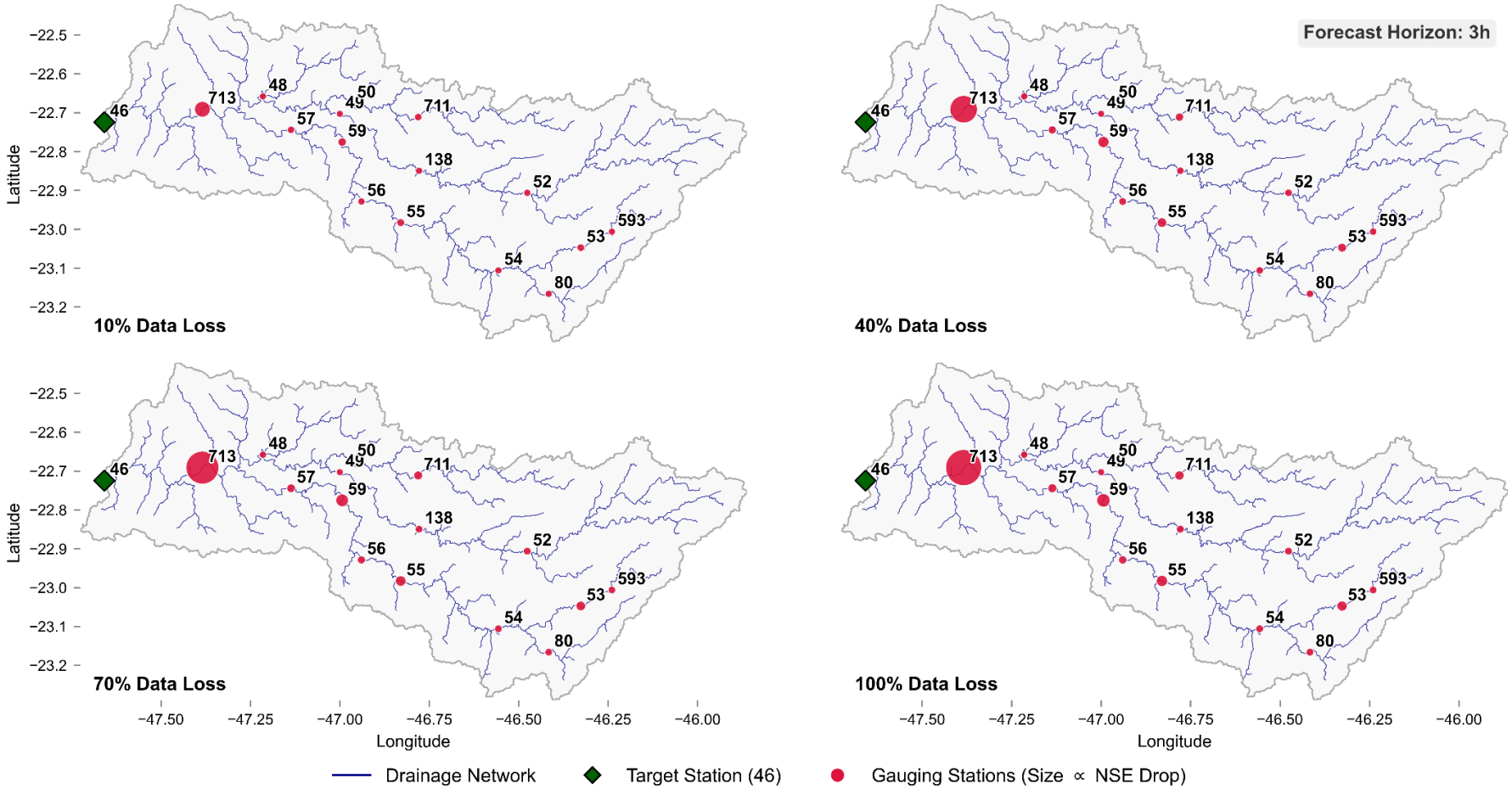


Figure 39C –Spatial representation of model performance (NSE) for 3 hours lead time with variable data loss.
Source: The author (2025).

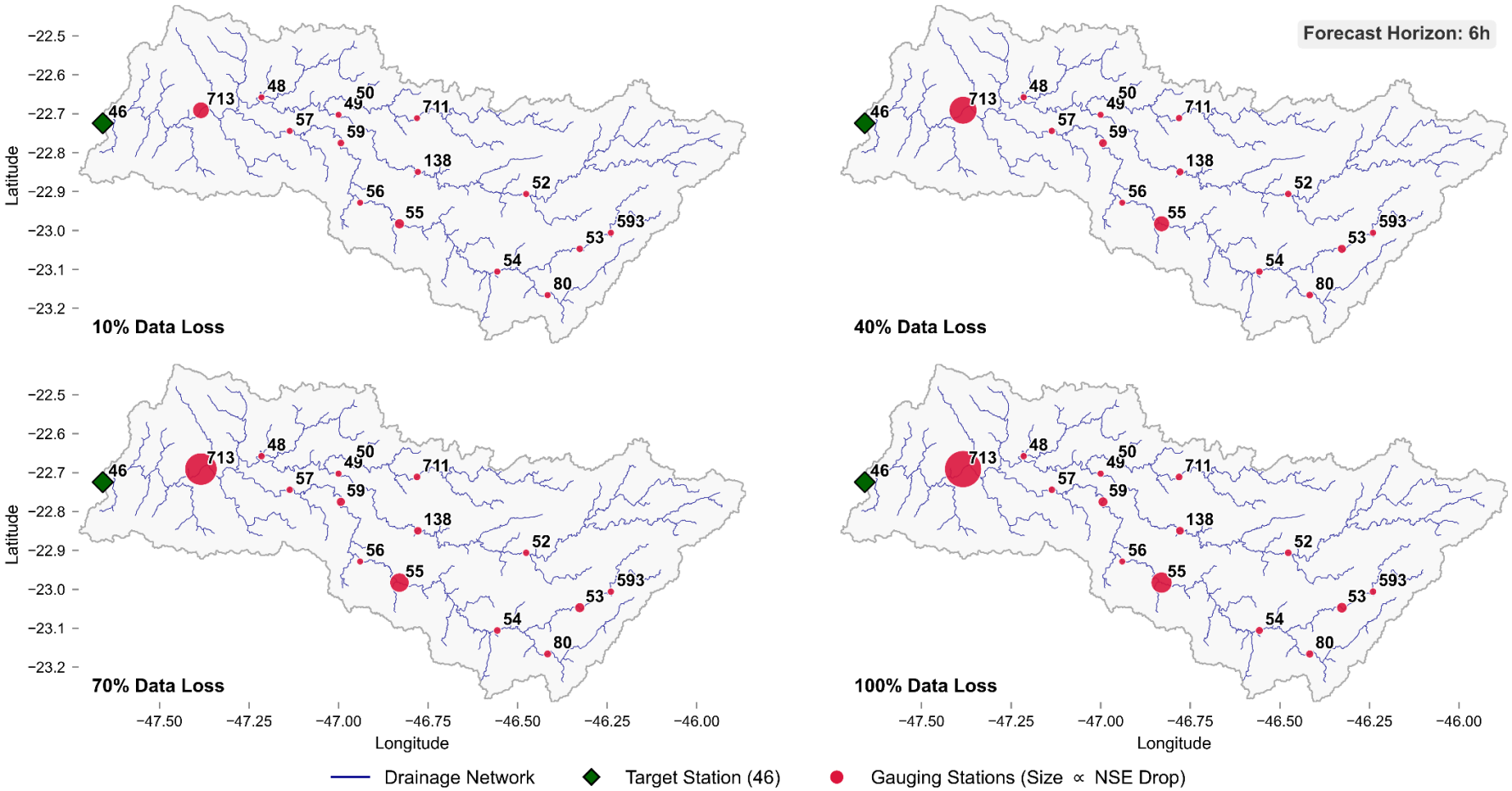


Figure 40C –Spatial representation of model performance (NSE) for 6 hours lead time with variable data loss.
Source: The author (2025).

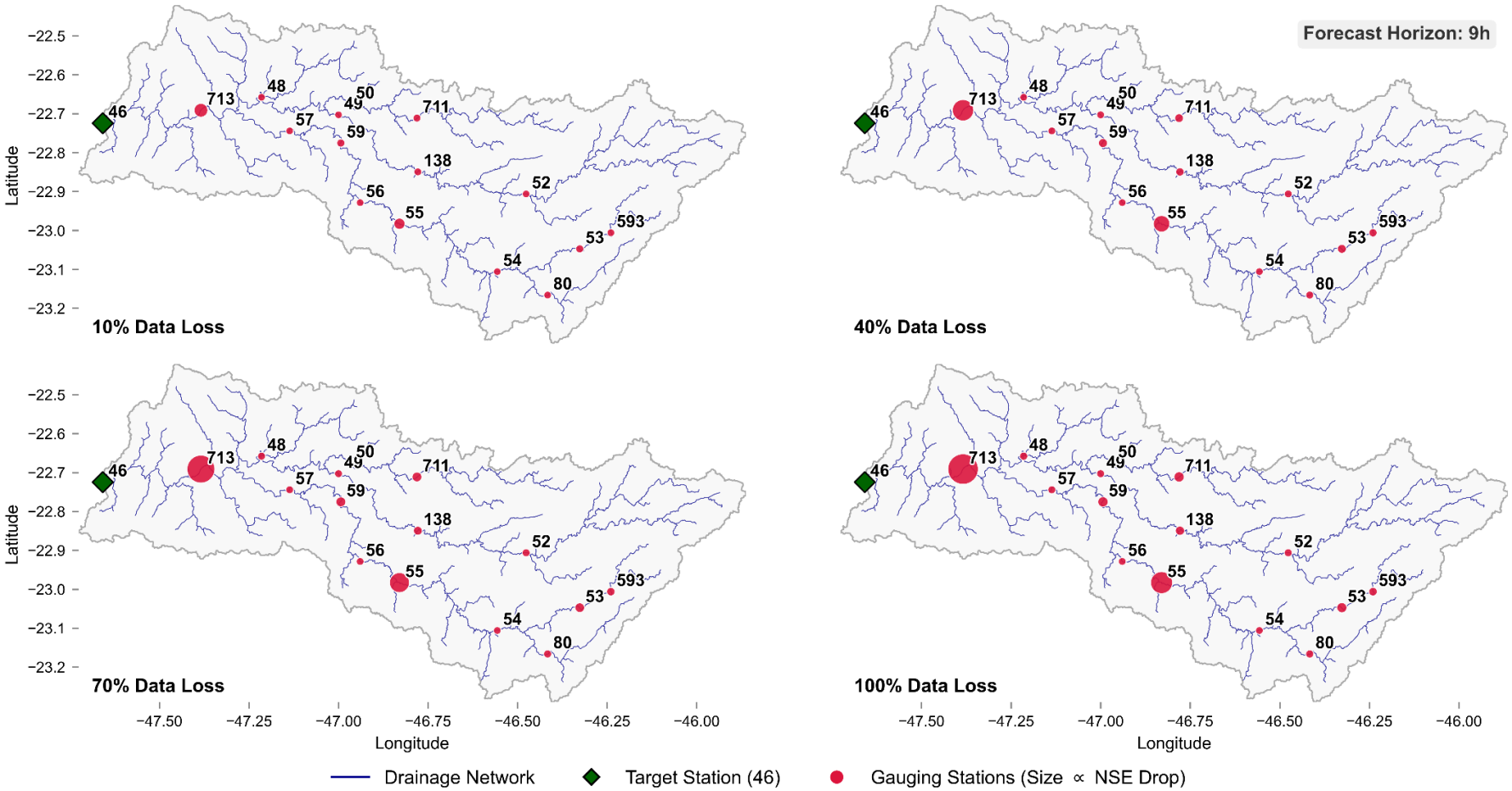


Figure 41C –Spatial representation of model performance (NSE) for 9 hours lead time with variable data loss.
Source: The author (2025).

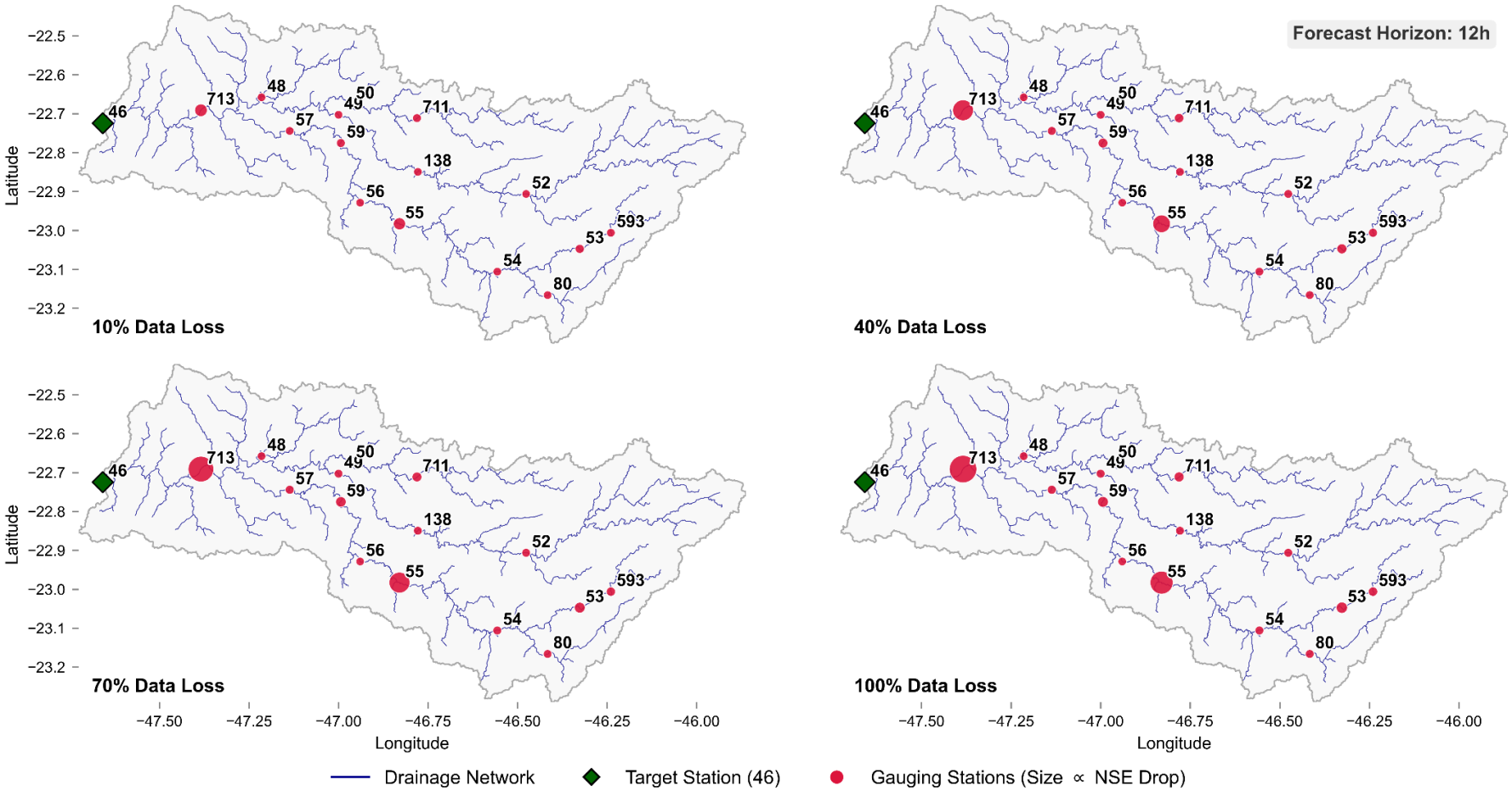


Figure 42C –Spatial representation of model performance (NSE) for 12 hours lead time with variable data loss.
Source: The author (2025).

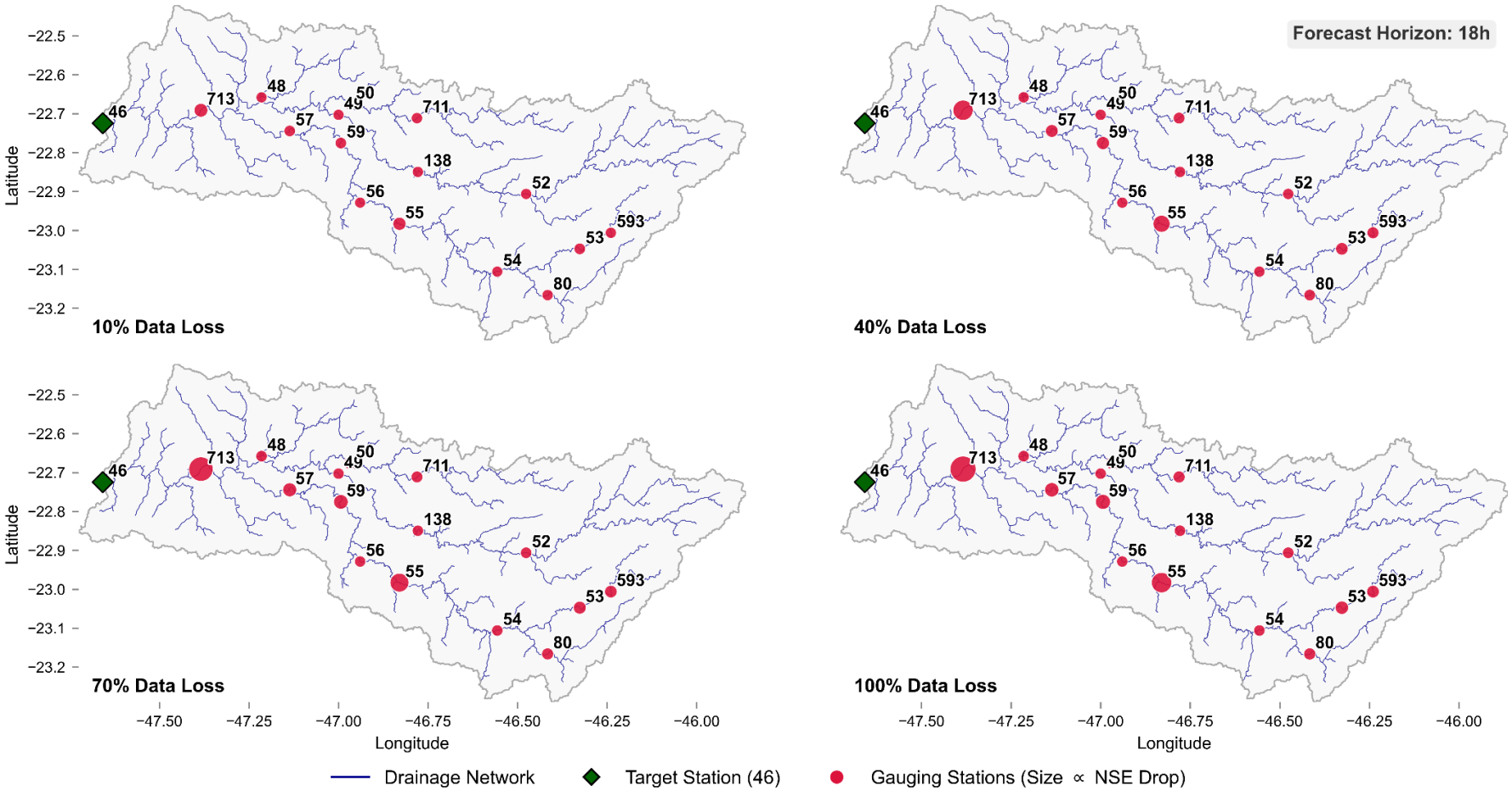


Figure 43C –Spatial representation of model performance (NSE) for 18 hours lead time with variable data loss.
Source: The author (2025).

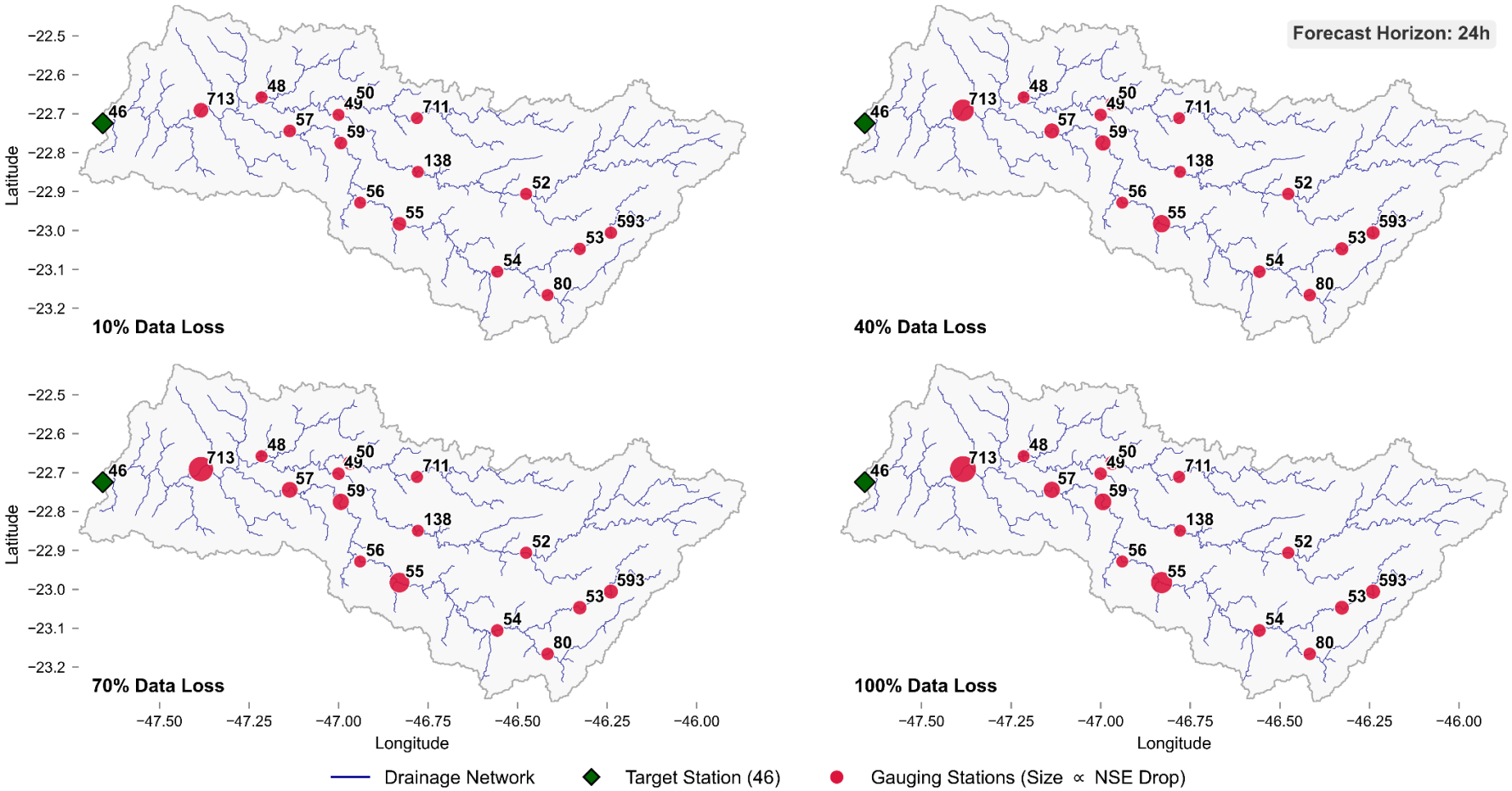


Figure 44C –Spatial representation of model performance (NSE) for 24 hours lead time with variable data loss.
Source: The author (2025).

APPENDIX D – SHAP Gauge Importance Rank

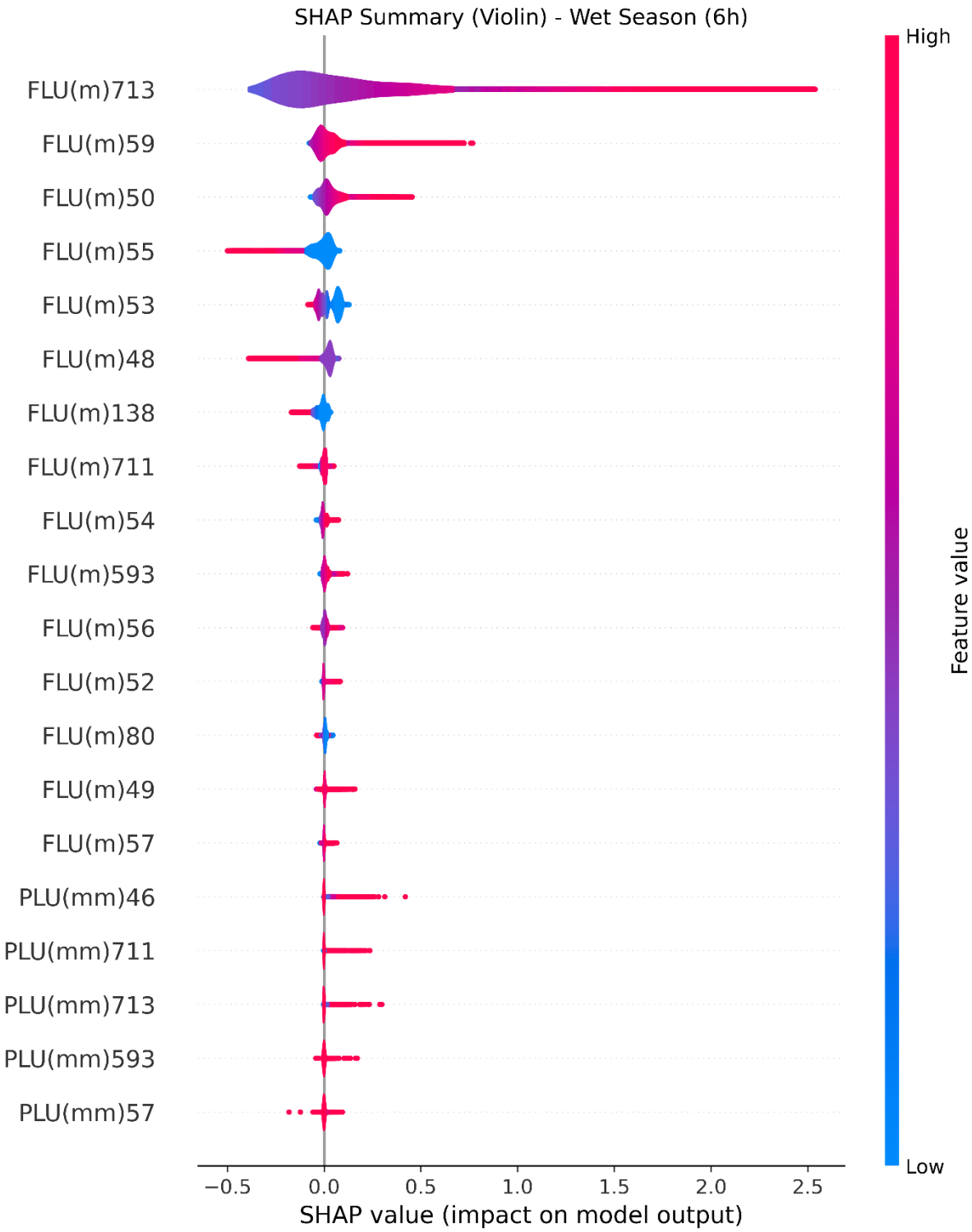


Figure 45D – SHAP Values gauge-ranking for 6-hour forecast

Source: The author (2025).

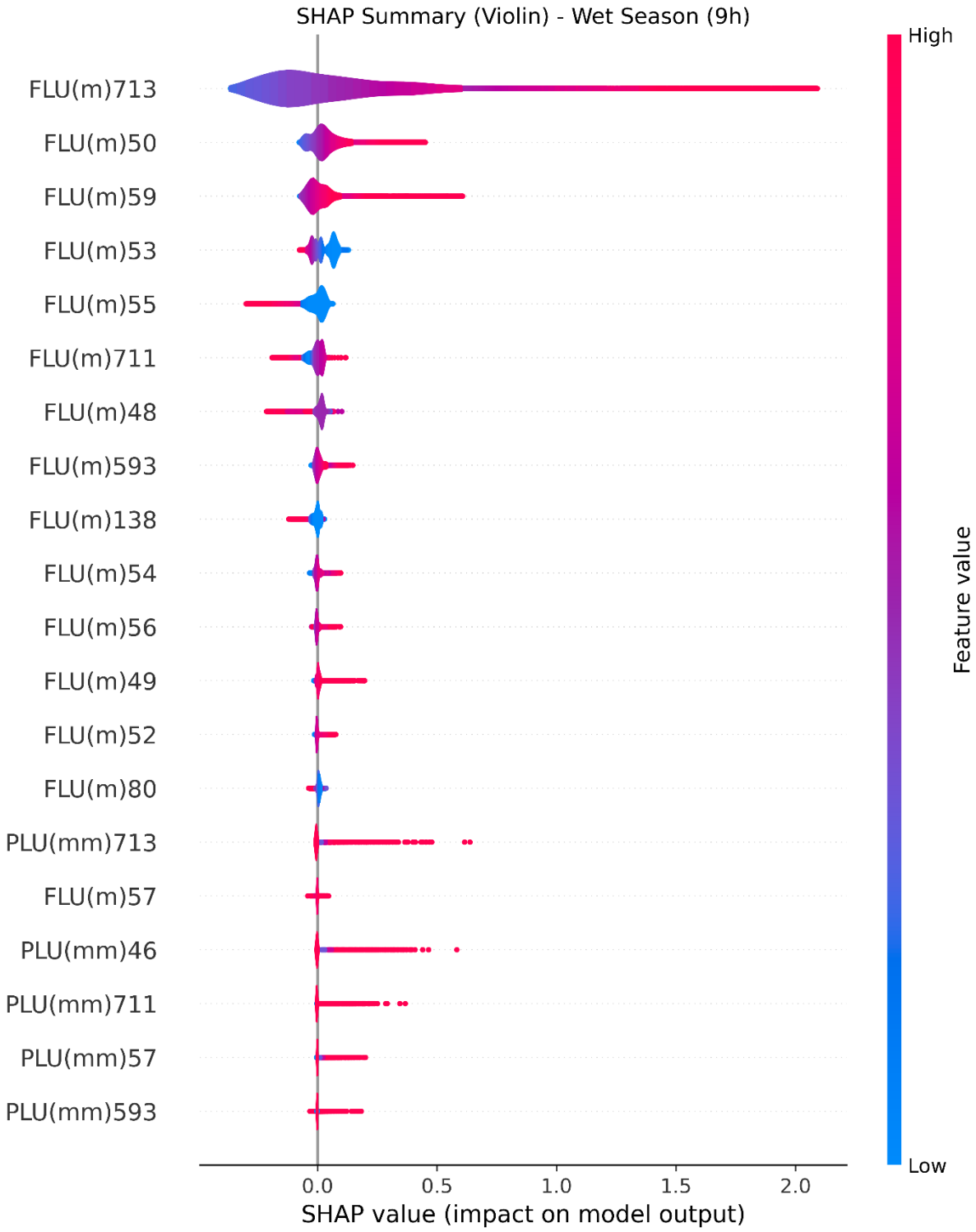


Figure 46D – SHAP Values gauge-ranking for 9-hour forecast

Source: The author (2025).

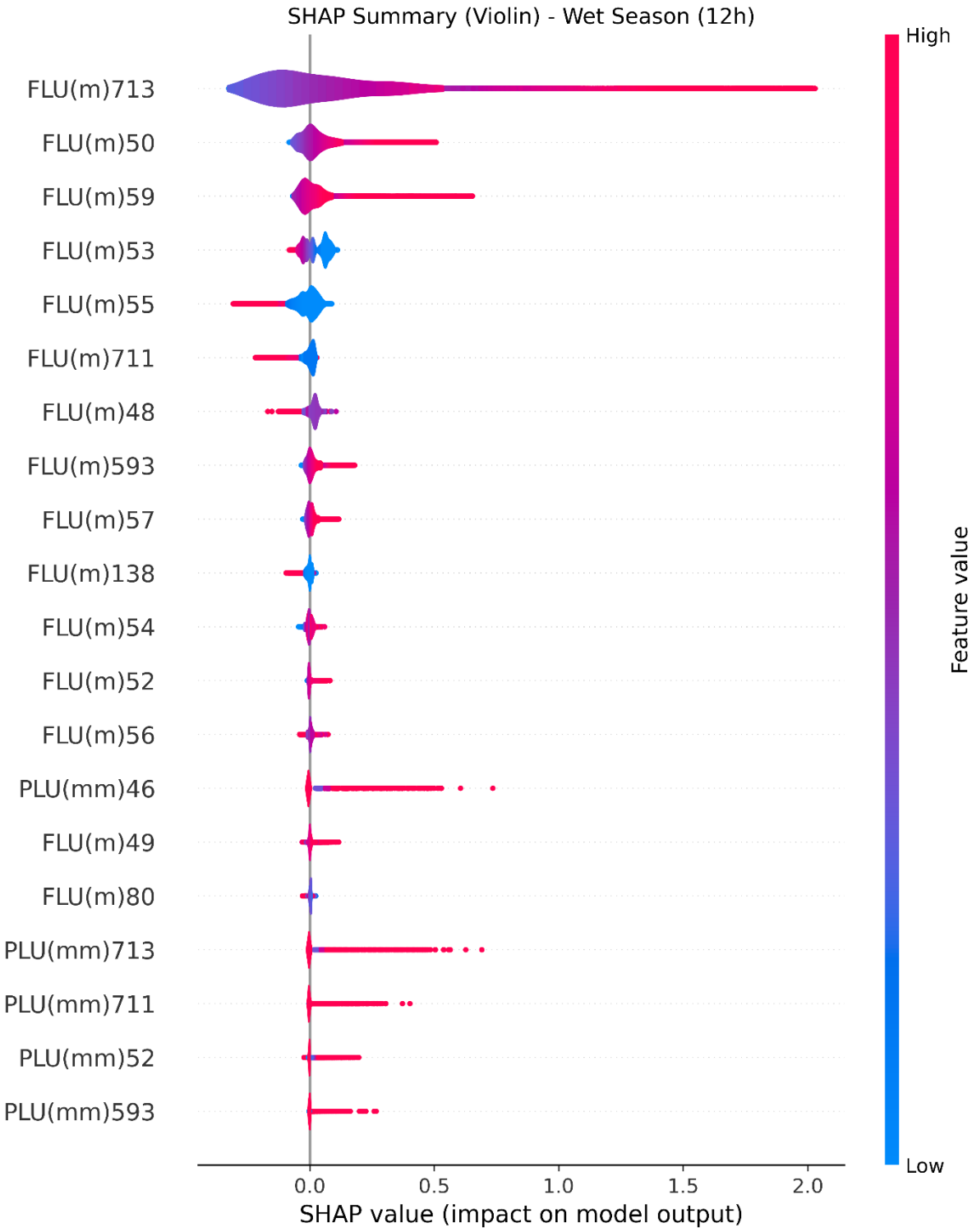


Figure 47D – SHAP Values gauge-ranking for 12-hour forecast

Source: The author (2025).

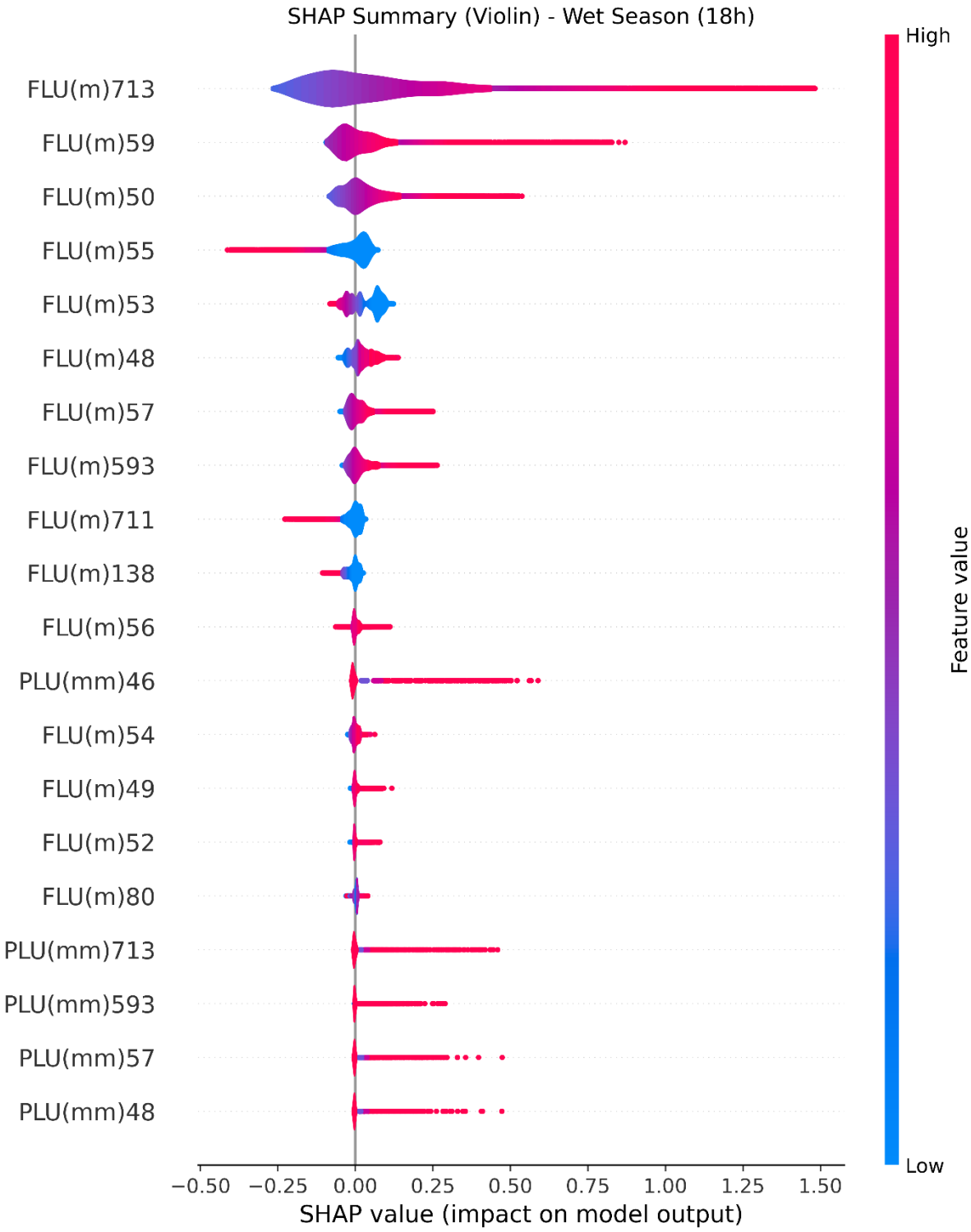


Figure 48D – SHAP Values gauge-ranking for 18-hour forecast

Source: The author (2025).

APPENDIX E – SHAP Gauge Importance Maps

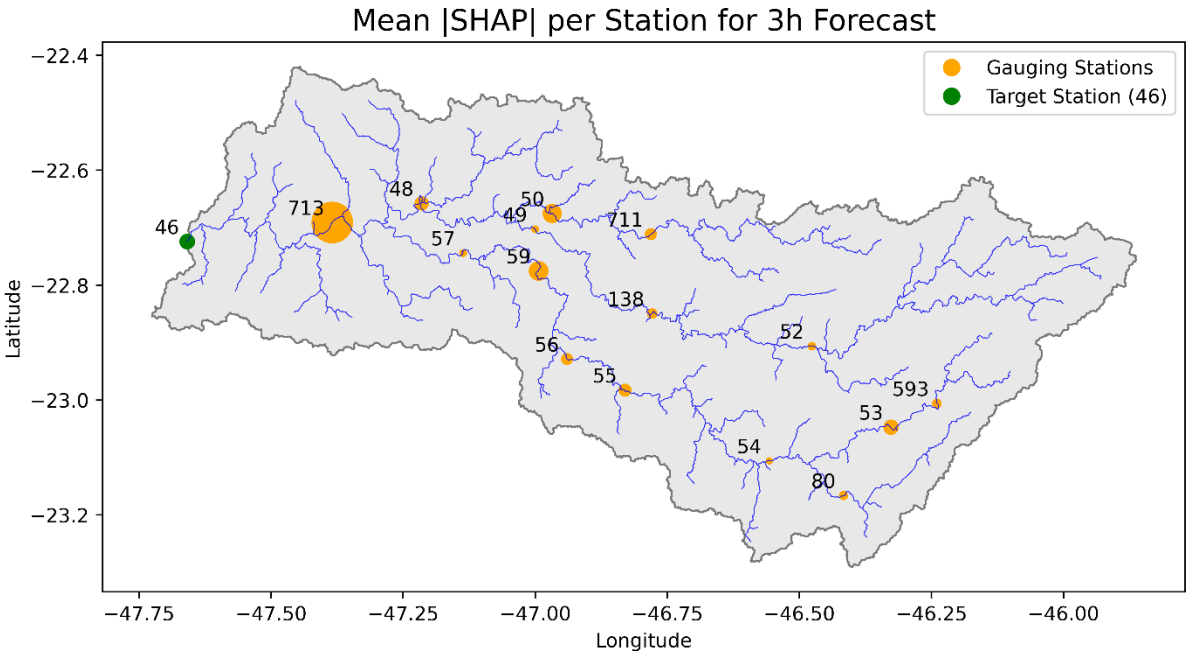


Figure 49E – SHAP Values Importance Map for 3-hour forecast.
Source: The author (2025).

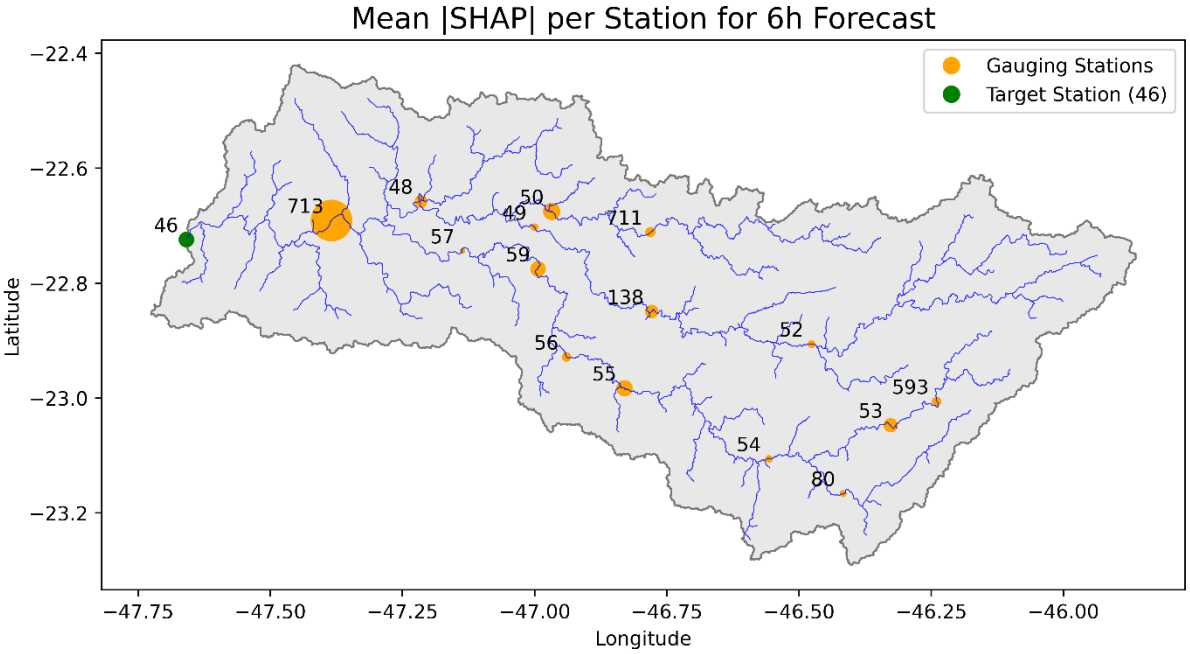


Figure 50E – SHAP Values Importance Map for 6-hour forecast.
Source: The author (2025).

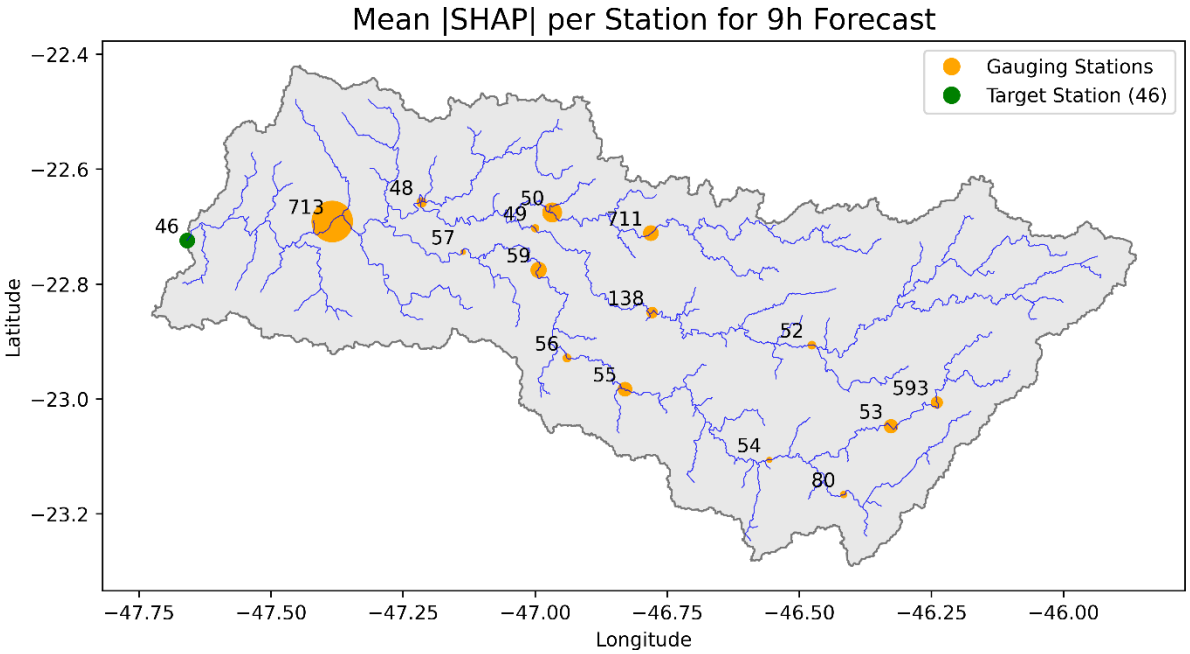


Figure 51E – SHAP Values Importance Map for 9-hour forecast.
Source: The author (2025).

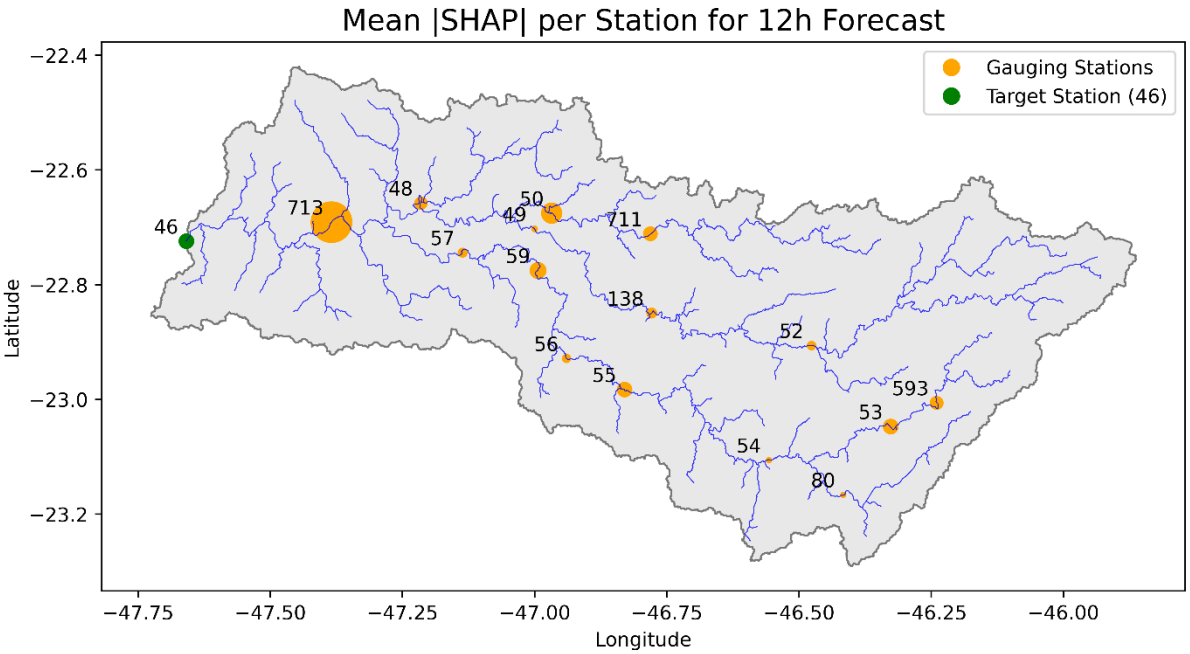


Figure 52E – SHAP Values Importance Map for 12-hour forecast.
Source: The author (2025).

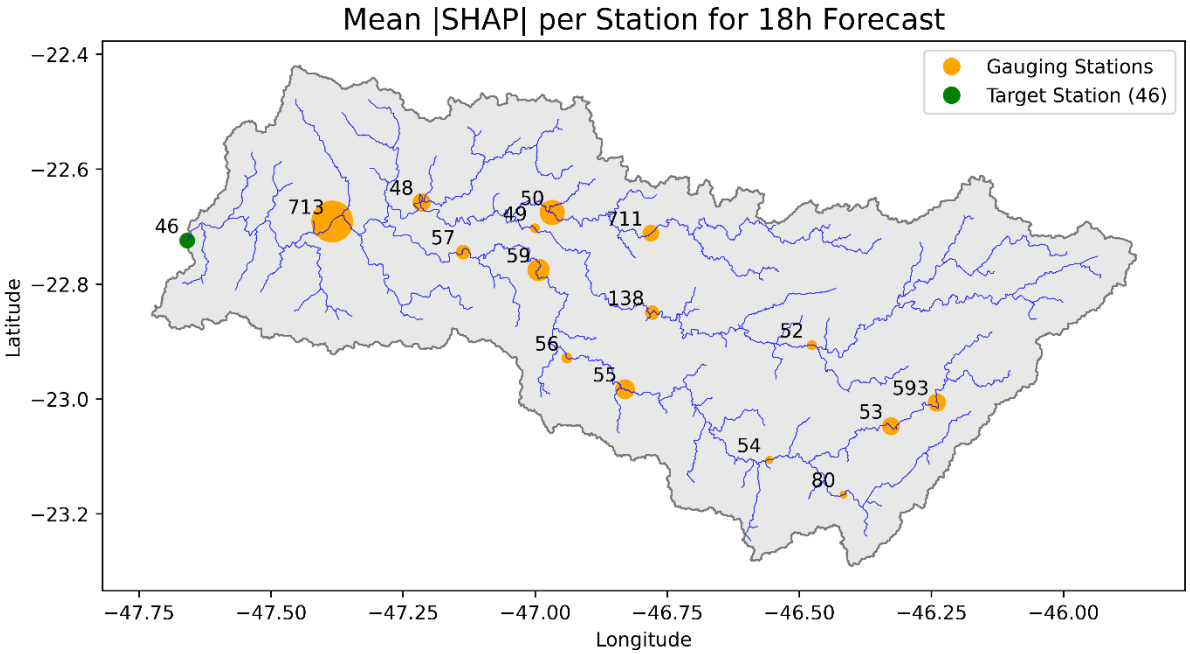


Figure 53E – SHAP Values Importance Map for 18-hour forecast.
Source: The author (2025).

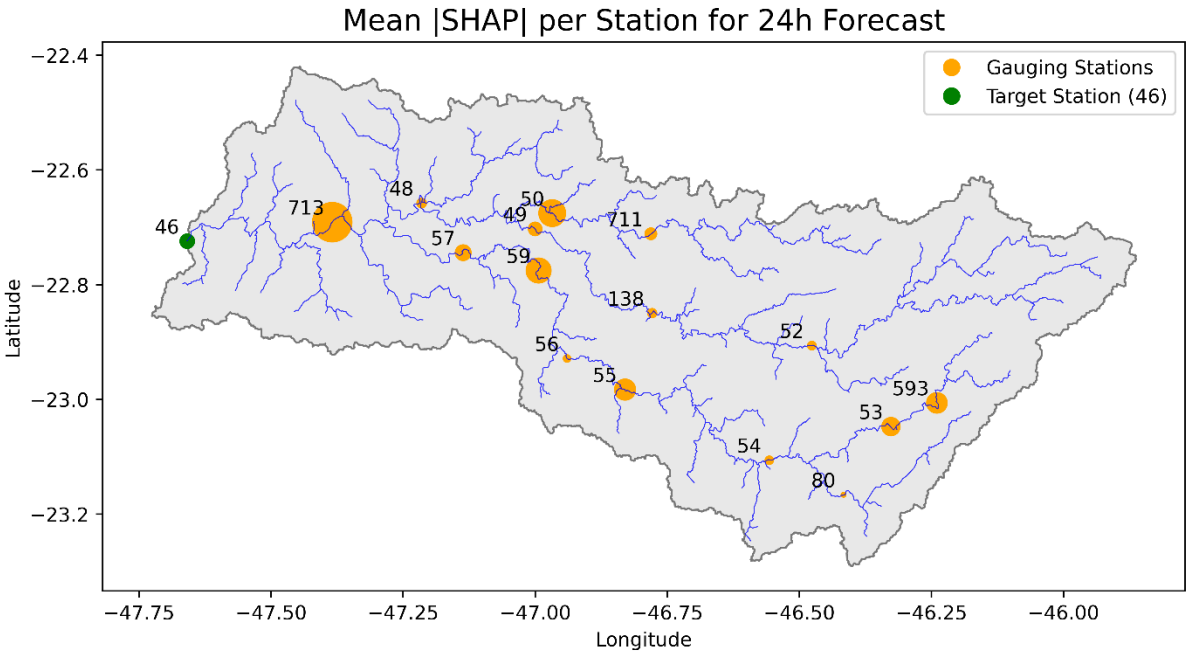


Figure 54E – SHAP Values Importance Map for 24-hour forecast.
Source: The author (2025).

APPENDIX F – SHAP Values Over Time

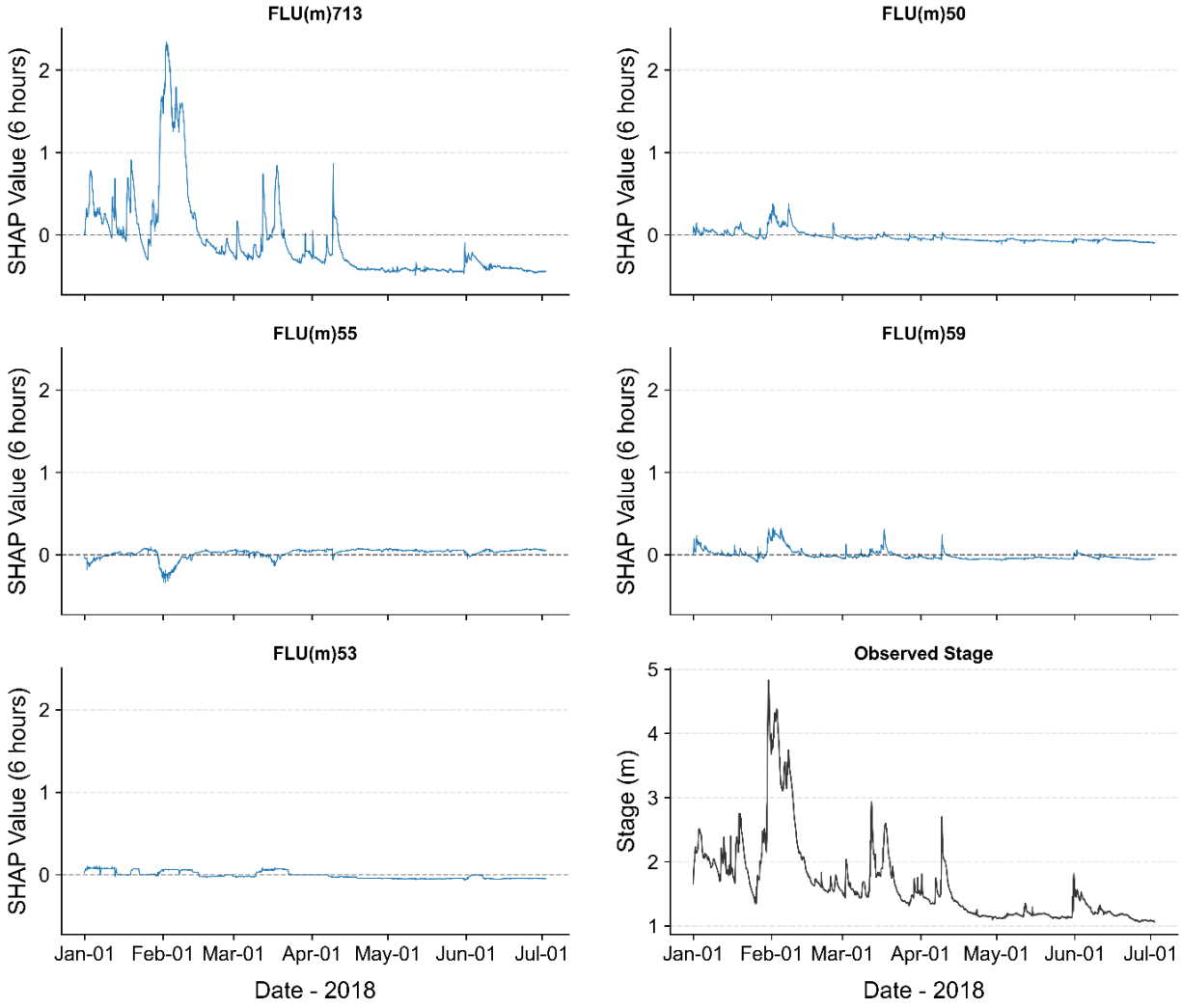


Figure 55F – SHAP Values over time for 6-hour lead time.

Source: The author (2025).

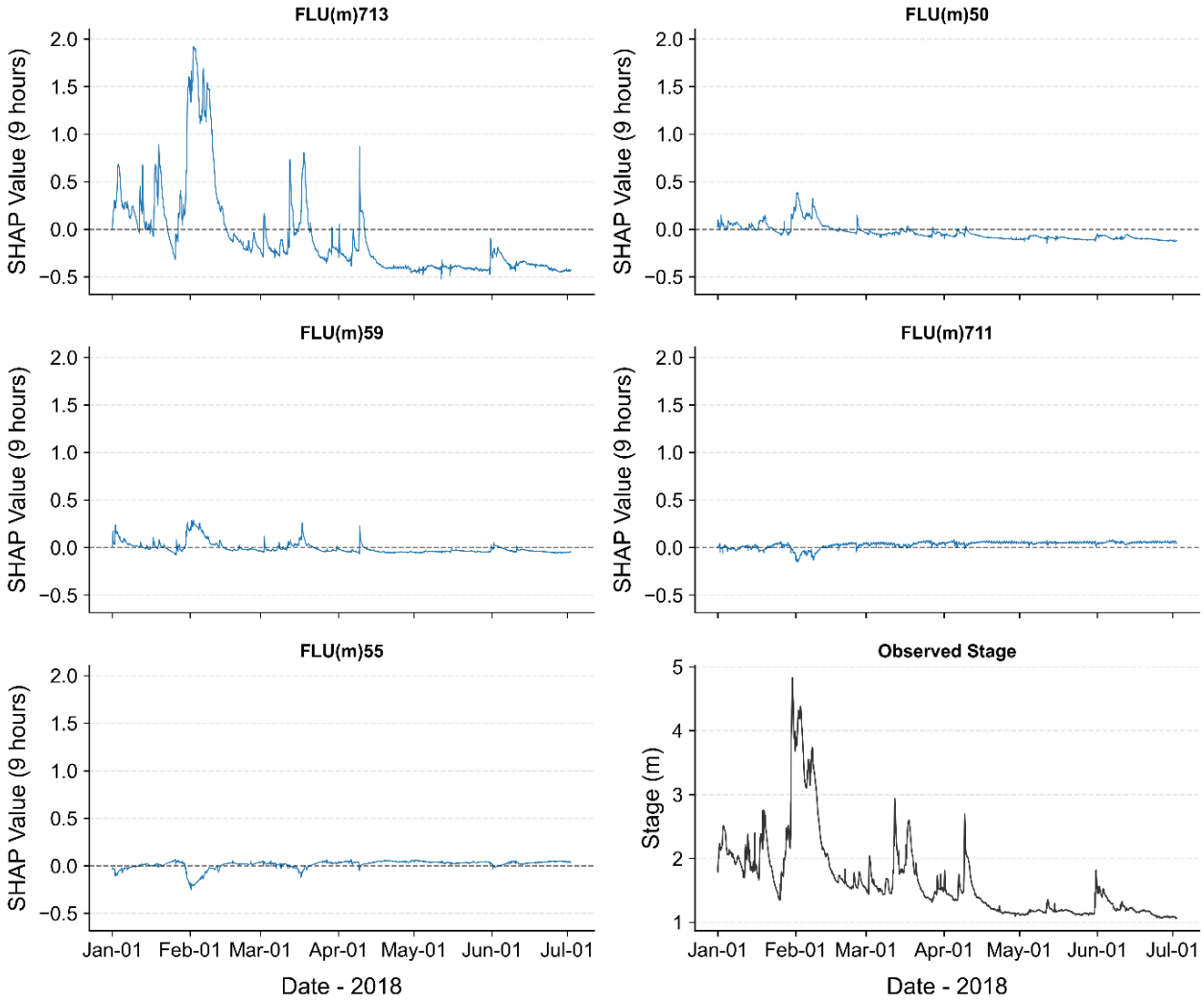


Figure 56F – SHAP Values over time for 9-hour lead time.

Source: The author (2025).

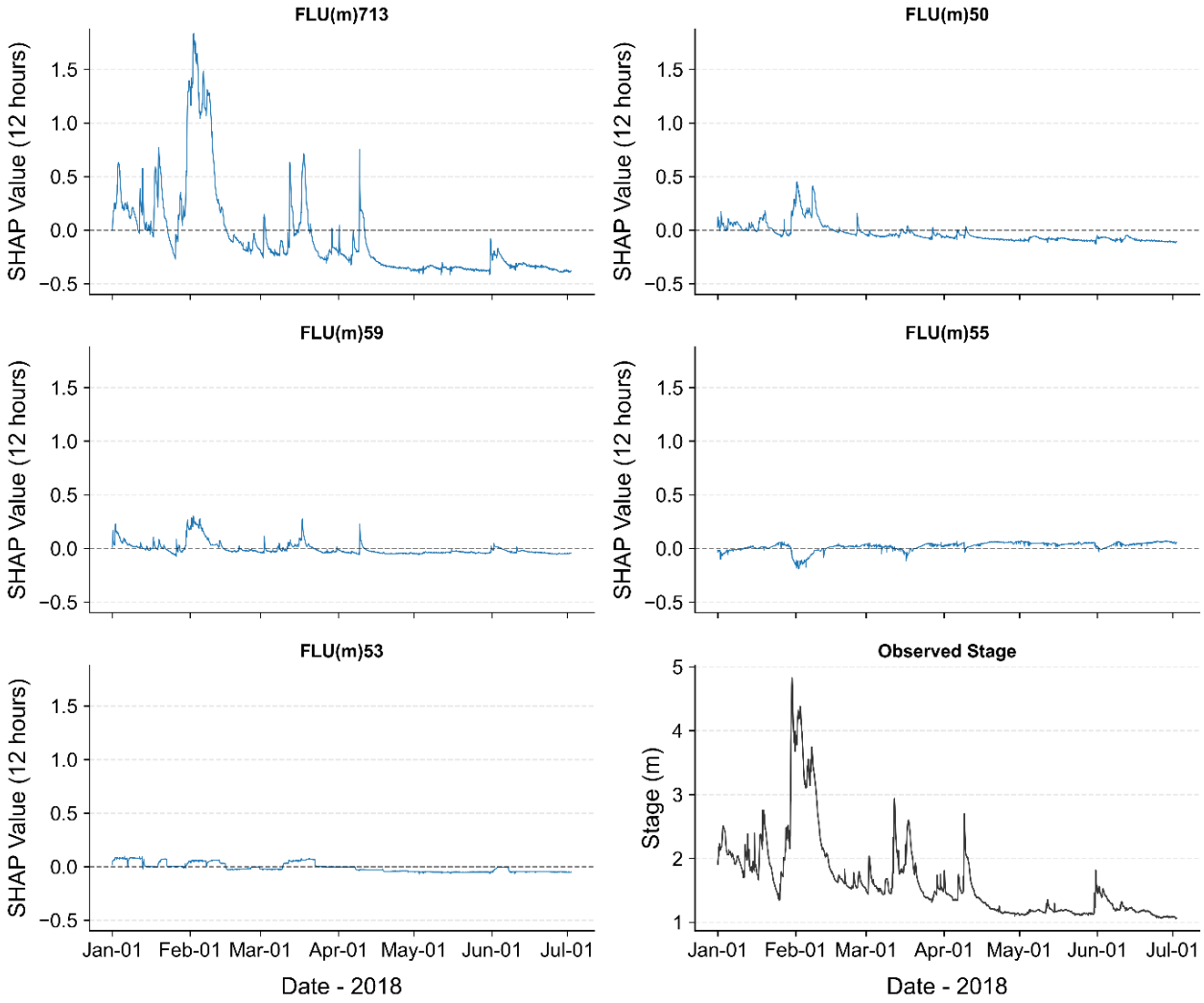


Figure 57F – SHAP Values over time for 12-hour lead time.
Source: The author (2025).

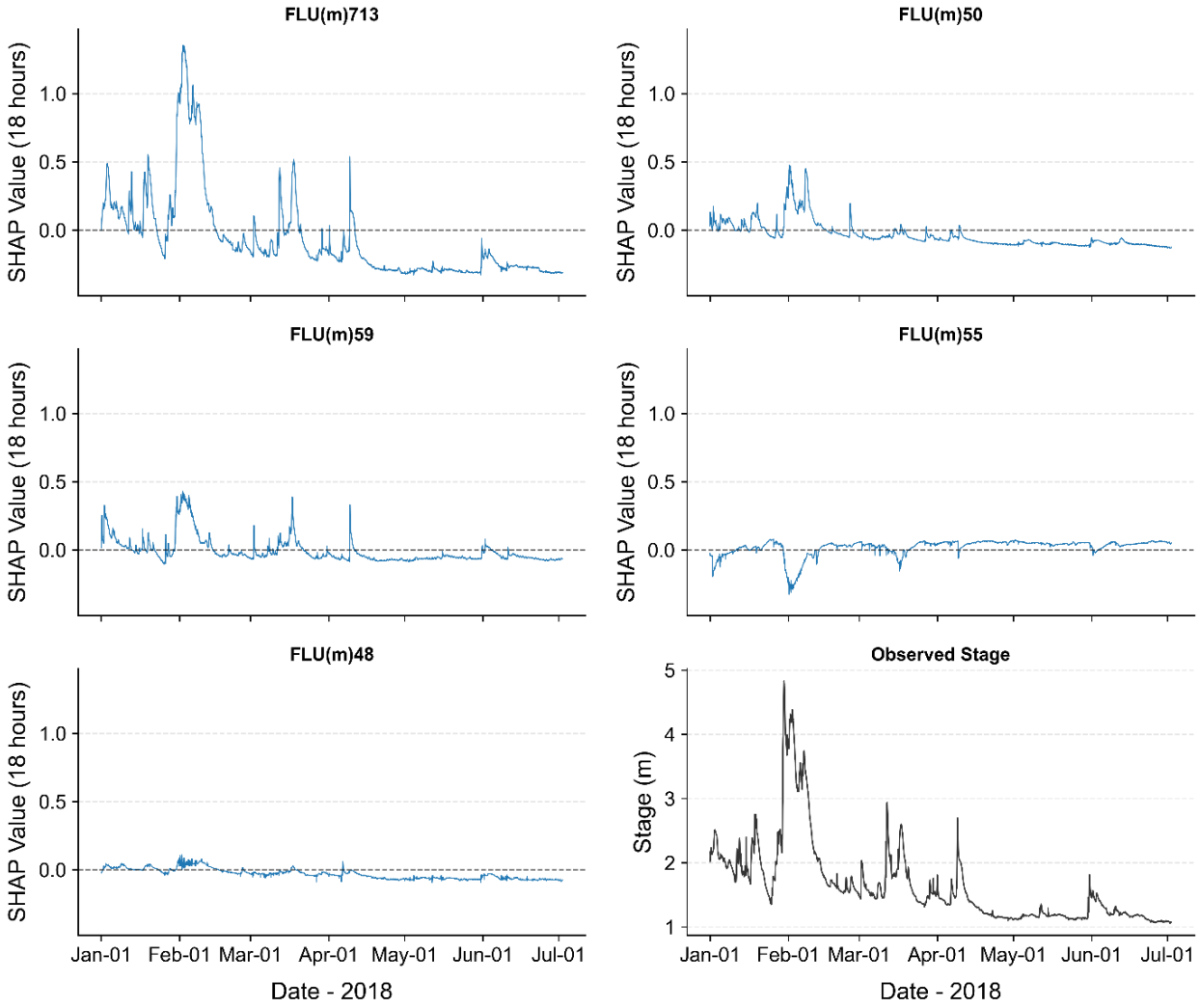


Figure 58F – SHAP Values over time for 18-hour lead time.
Source: The author (2025).

APPENDIX G – Prioritized Network Data-Deterioration Maps

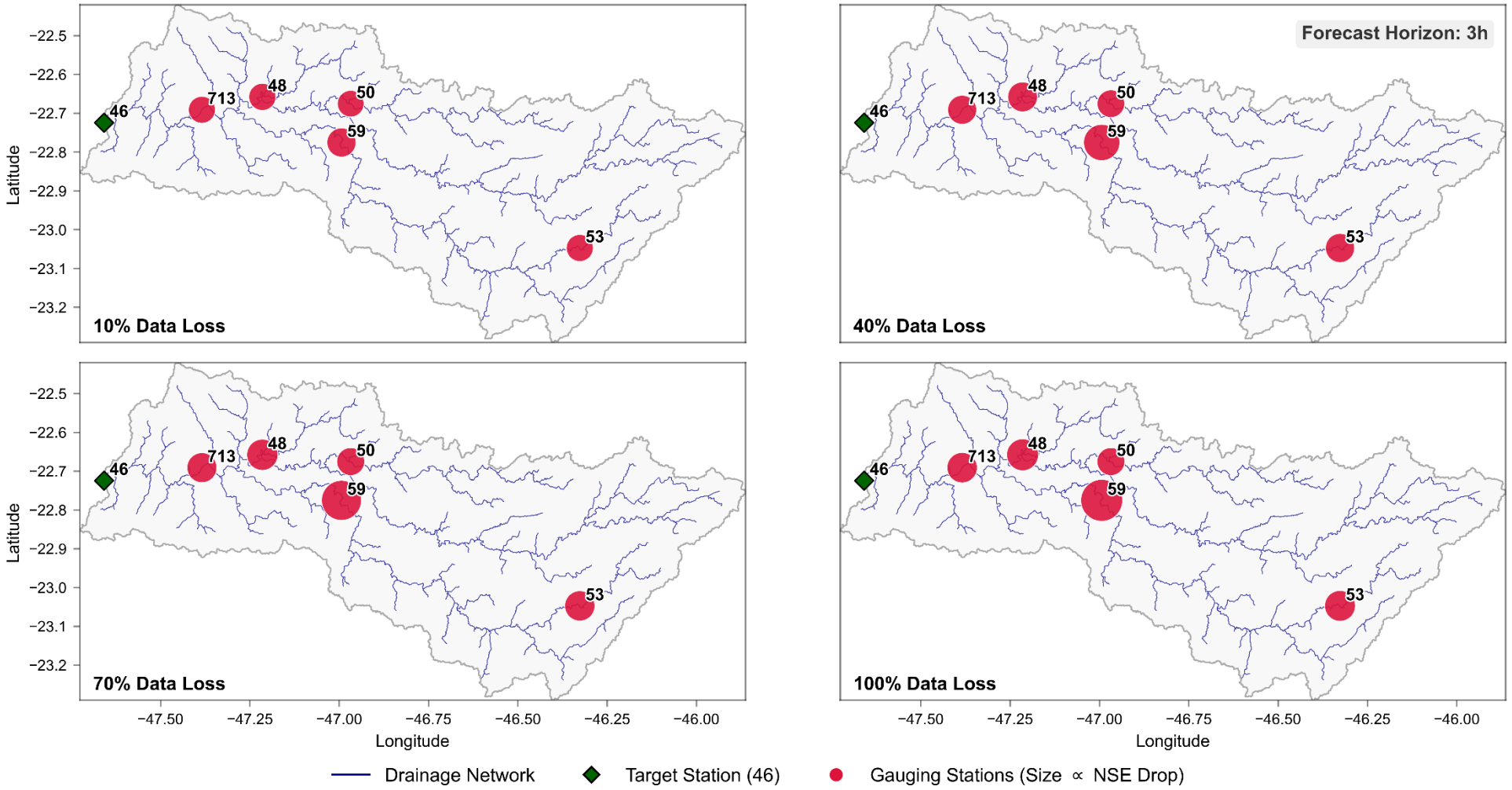


Figure 59E – Spatial representation of model performance (NSE) for 3 hours lead time with variable data loss (Prioritized Network).
Source: The author (2025).

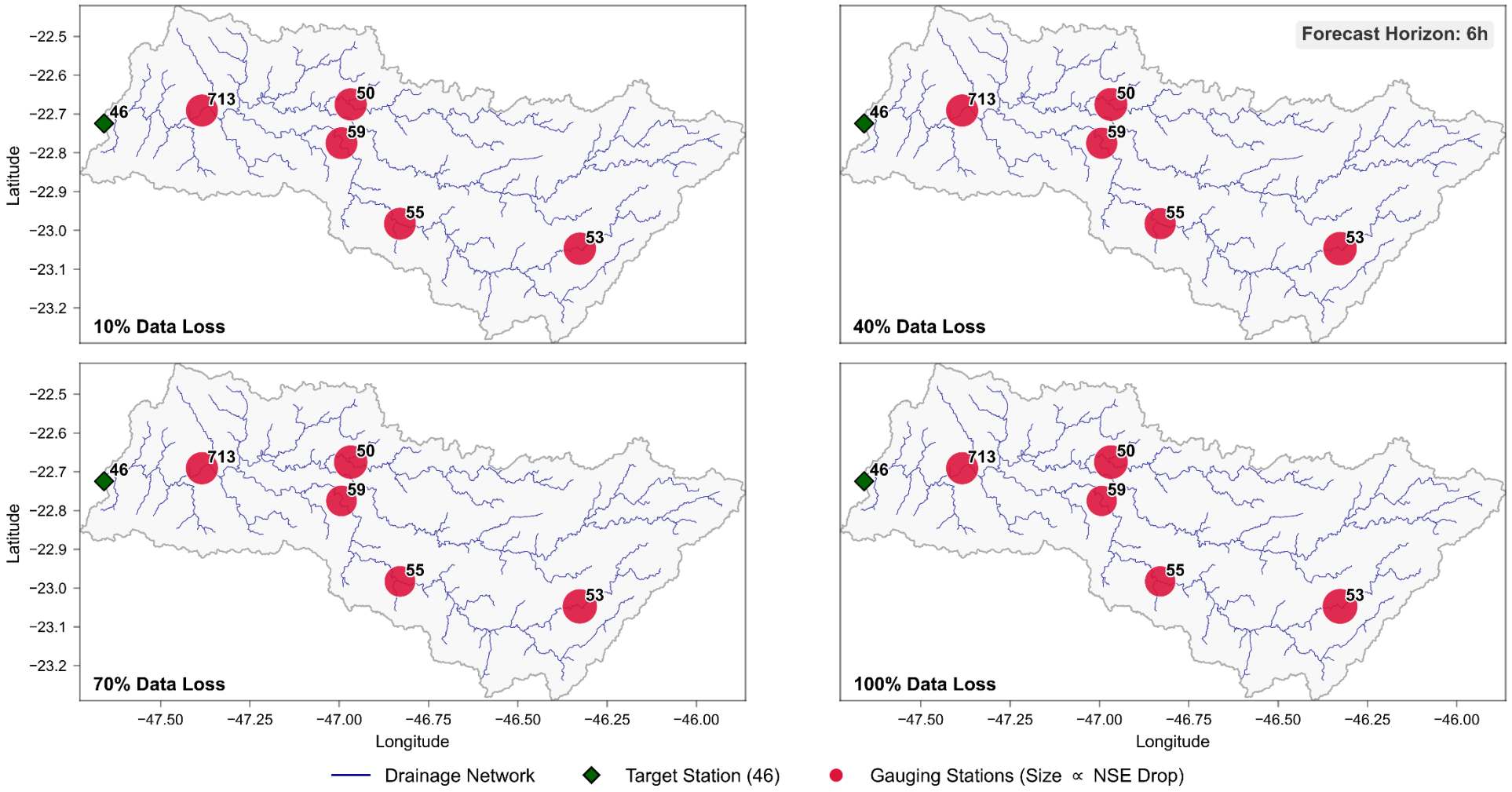


Figure 60E – Spatial representation of model performance (NSE) for 6 hours lead time with variable data loss (Prioritized Network).
Source: The author (2025).

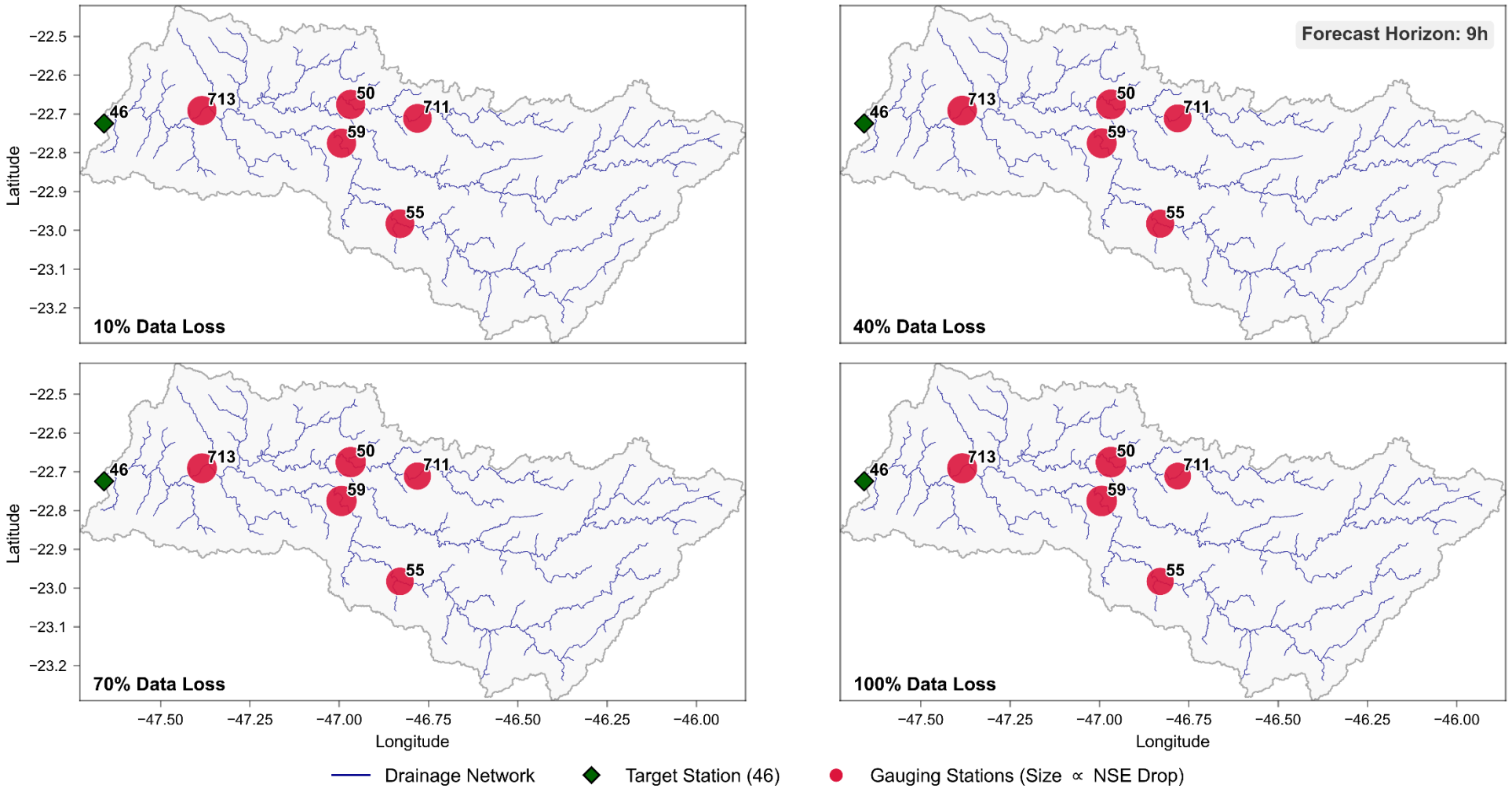


Figure 61E – Spatial representation of model performance (NSE) for 9 hours lead time with variable data loss (Prioritized Network).
Source: The author (2025).

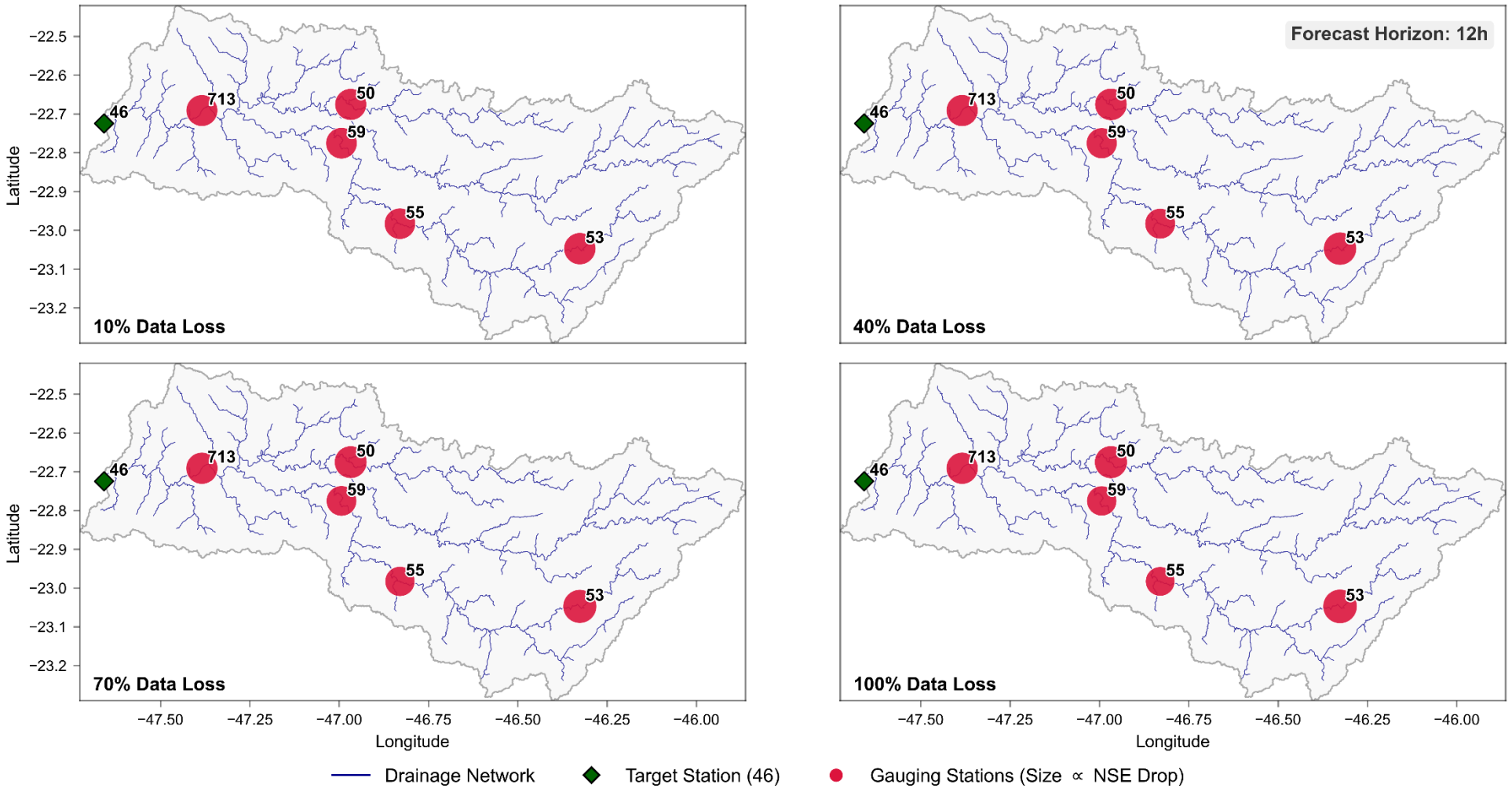


Figure 62E – Spatial representation of model performance (NSE) for 12 hours lead time with variable data loss (Prioritized Network).
Source: The author (2025).

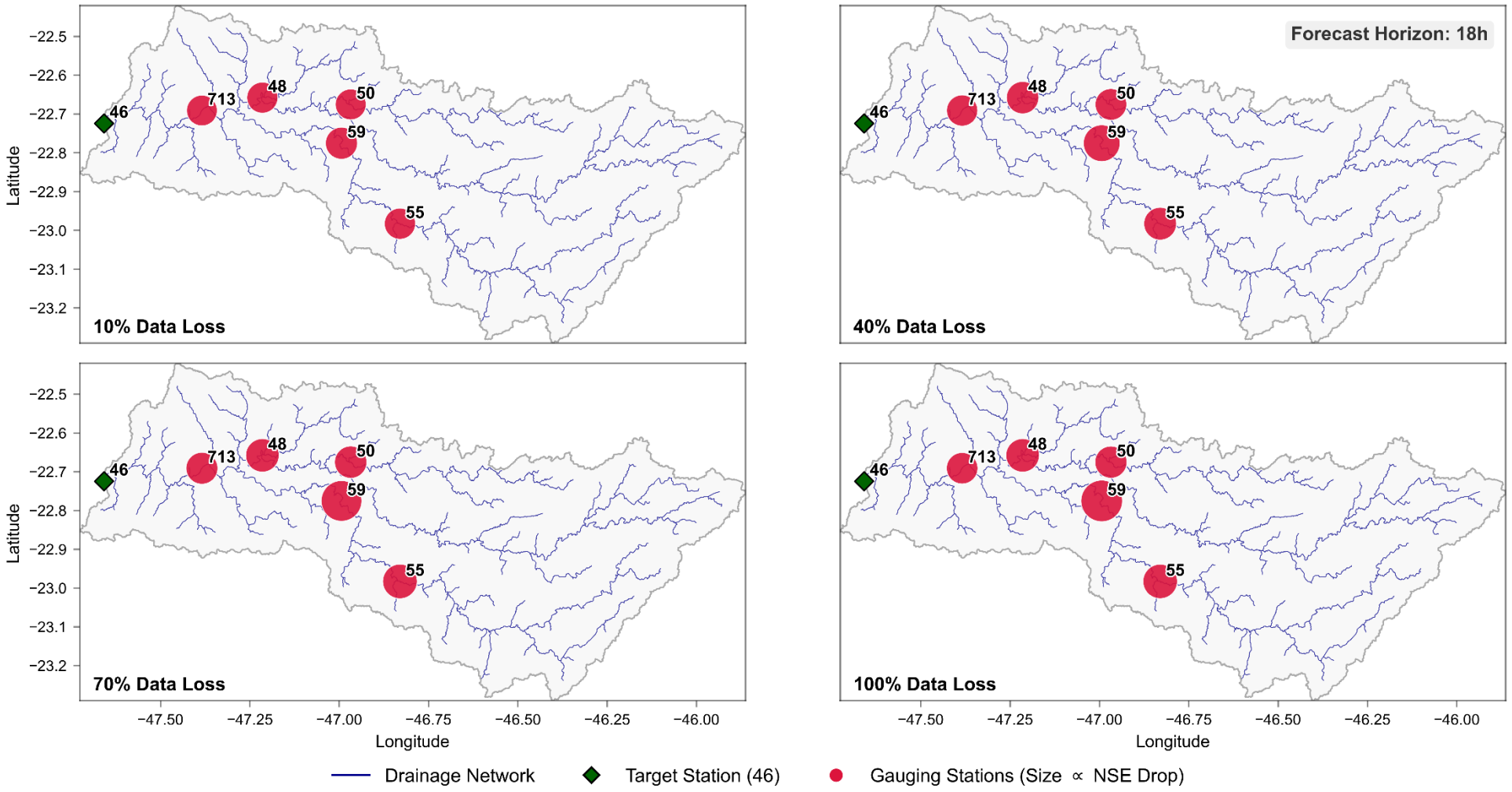


Figure 63E – Spatial representation of model performance (NSE) for 18 hours lead time with variable data loss (Prioritized Network).
Source: The author (2025).

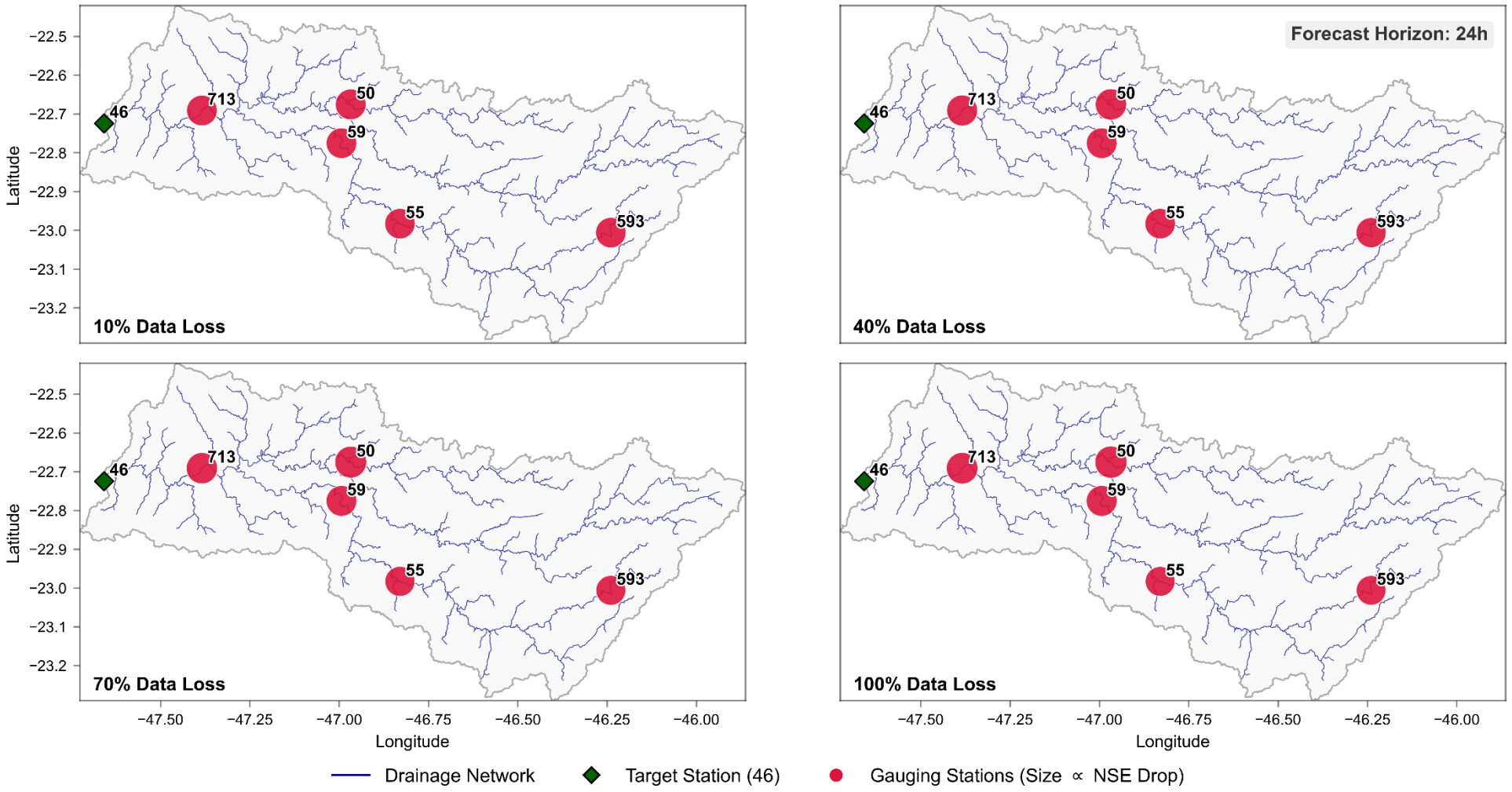


Figure 64E – Spatial representation of model performance (NSE) for 24 hours lead time with variable data loss (Prioritized Network).
Source: The author (2025).