applied sciences

MDPI

Article

Contextual Reuse of Big Data Systems: A Case Study Assessing Groundwater Recharge Influences

Agustina Buccella [1,*], Alejandra Cechich [1], Walter Garrido [1] and Ayelén Montenegro [2]

[1] Facultad de Informática, Universidad Nacional del Comahue, Neuquen 8300, Argentina; alejandra.cechich@fi.uncoma.edu.ar (A.C.); walter.garrido@est.fi.uncoma.edu.ar (W.G.)
[2] Instituto Nacional de Tecnología Agropecuaria (INTA), Alto Valle de Río Negro y Neuquén, Allen 8332, Río Negro, Argentina
* Correspondence: agustina.buccella@fi.uncoma.edu.ar

Abstract

The process of building data analytics systems, including big data systems, is currently being investigated from various perspectives that generally focus on specific aspects, such as data security or privacy, to the detriment of an engineering perspective on systems development. To address this limitation, our proposal focuses on developing analytics systems through a reuse-based approach, including stages ranging from problem definition to results analysis by identifying variations and building reusable, context-based assets. This study presents the reuse process by constructing two case studies that address the water table level prediction problem in two different contexts: the irrigated period and the non-irrigated period in the same study area. The objective of this study is to demonstrate the influence of context on the performance of widely used predictive models for this problem, including long short-term memory (LSTM), artificial neural networks (ANNs), and support vector machines (SVMs), as well as the potential for reusing the developed analytics system. Additionally, we applied the permutation feature importance (PFI) to determine the contribution of individual variables to the prediction. The results confirm that the same problem hypotheses yield different performance in each case in terms of coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), and mean square error (MSE). They also show that the best-performing predictive models differ for some of the hypotheses (ANN in one case and LSTM in another), supporting the assumption that context can influence model selection and performance. Reusing assets allows for more efficient evaluation of these alternatives during development time, resulting in analytics systems that are more closely aligned with reality, while also offering the advantages of software system composition.

Keywords: big data systems; context reusability; precision agriculture; prediction grounwater levels

Academic Editor: José Miguel Molina Martínez

Received: 29 December 2025
Revised: 30 January 2026
Accepted: 4 February 2026
Published: 6 February 2026

1. Introduction

Nowadays, organizations are adopting a differentiated approach to data usage, going further than simple storage and querying. The real challenge now lies in taking advantage of the information generated, particularly given the vast amount of information available across various formats and systems. Much of these data remains unused, wasting its potential to generate strategic value. As a result, new types of requirements are emerging that focus primarily on the rapid and efficient utilization of all available information. Managing these new requirements is now possible thanks to the big data area, which

includes techniques, methods, and technologies needed to address specific challenges, such as the storage and management of large volumes of data, the velocity at which data are generated, the variety of data formats and sources, etc. [1,2].

In particular, within the field of software engineering and intelligent systems, methodologies are emerging for the development of big data systems (BDSs) that provide the software, techniques, and resources necessary to collect, transform, analyze, and visualize data, with a particular focus on supporting decision-making. These new methodologies [3–6] identify the development of BDSs from three main perspectives:multidisciplinary, domain-oriented, and software. The first highlights the involvement in the development process of actors from various disciplines, including computer scientists, technicians, and domain experts. The second emphasizes the domain knowledge in which the BDS is situated, allowing a better understanding, guidance, and improvement of the activities and outcomes. Finally, the third considers domain-specific characteristics to produce common and variant artifacts to support the development of these systems according to the principles of reuse and flexibility.

With respect to the domain-oriented perspective, our focus is on contextual reuse, meaning that the choice of analytic features and models depends not only on the nature of the problem but also on the environment under study. In our experience, models that perform well in certain environments do not necessarily work the same way in others. However, the same problem addressed in different environments has common characteristics and variations, which allows certain aspects to be reused. Based on these assumptions, contextual reuse allows us to identify variations in case analysis that may stem from data sources, selected variables, or the analytics themselves. All these variations can be documented and form the basis for building systems better suited to the specific context.

To highlight this point, herein, we present an extension of the study presented in [7], in which we developed two application cases within the precision agriculture domain. Specifically, we identified and verified hypotheses from expert users focused on the requirement of *analyzing the influences of weather and river flow variables on the behavior of the water table in the region*. The hypotheses were evaluated in two different application contexts: (1) during farm irrigation periods and (2) during non-irrigation periods. The extension proposed here aims to continue working with the same requirement but adds new hypotheses defined by expert users to predict groundwater recharge by considering influencing factors. To do this, we apply our reuse methodology, showing how previous application cases can be effectively reused to support the new hypotheses.

Thus, the main contributions of this paper are summarized as follows: (1) a description of the reuse process activities instantiated in two case studies at different periods (irrigation and non-irrigation) and three associated hypotheses; and (2) a comparative analysis of the performance of the applied predictive models showing differences in best performance (e.g., ANN for one period and LSTM [8] for the other considering the same hypothesis) for the other considering the same hypothesis), which would indicate an influence of the context variety and consequently, of other associated varieties, some of them reusable. These contributions can help improve the information available for decision-making regarding starting/ending irrigation periods, as well as controlling the amount of water needed in a variable and environment-dependent manner.

This article is organized as follows: In the next section, we summarize related work on the reuse of data analytics systems, highlighting the differences with our approach, which is briefly presented. The subsequent section describes the materials and methods, introducing the domain of the case studies, along with related work on groundwater recharge analysis. Our previous work in this area is also discussed to showcase existing assets that predate the studies presented in this article. Section 4 specifically describes our process for developing

contextual reuse cases by outlining activity diagrams in BPMN (Business Process Modeling Language), which are instantiated in the new case studies. A discussion, conclusions, and future work are addressed at the end.

2. Related Work and Background

In BDSs, reusability has been approached from various angles. For example, in [9], some concepts are discussed in the context of data analytics distinguishing between data use and reuse. Various open research questions on reuse are proposed, such as tradeoffs between collecting new data and reusing existing ones; the need to distinguish between use and reuse, etc. More specifically, the work presented in [10] provides a deeper analysis of privacy considerations within the context of data reusability. Here, a data reuse taxonomy is proposed, which may be useful in determining to what extent that reuse should be allowed and under what conditions to preserve privacy. Other proposals address reuse issues in terms of increasing collaboration in the development of BDSs through the use of new technologies (i.e., cloud computing). For example, reference [11] proposes a management approach for BDSs by using storage and processing capabilities of a public cloud. Additionally, the different support platforms for the development of BDSs are also approached from the point of view of reuse; more specifically, the work in [5] analyzes the improvement in the efficiency of tools, such as Apache Hadoop (https://hadoop.apache.org/ (accessed on 1 February 2026)) and Spark (https://spark.apache.org/ (accessed on 1 February 2026)) due to the reuse of artifacts among different projects. To do so, common aspects are analyzed to provide a workflow implemented in a scalable and extensible way.

On the reuse management side, in [6], an exploratory analysis is carried out through interviews with Microsoft scientists, to collect information about which tasks within the life cycle are reused and, along with them, strategies for sharing and reusing previous works.

In addition to managing reuse across different teams, reusability was also addressed in terms of how architectures can be composed. In this sense, the work in [12] incorporates the detection of common and variable aspects within the development of BDSs as families of systems. This work presents a reference architecture that allows system designers to (1) define requirements—the reference architecture identifies significant requirements and shows variations depending on the type of requirement; (2) develop and evaluate solutions—the architecture identifies modules that must be developed in order to enable certain required capabilities; and (3) integrate systems—existing systems can be mapped to modules of the reference architecture, resulting in easy identification of points of conflict where interoperability between systems must be addressed. Components of the Component Off-The-Shelf (COTS) paradigm, or other reusable technologies, can be mapped to particular modules within this architecture, which allows for the evaluation of how different technologies may contribute to the development of the solution.

There are some reference architectures for BDSs that propose the addition of semantics to their components to expand domain and design knowledge. For example, the work in [4] compares different architectures, concluding that three proposals (Bolster, Solid, and Polystore) add semantics so that the underlying data schema is understood by a machine; that is, when describing data and their characteristics/relations. For instance, Bolster [13] adds a semantic layer that contains a metadata management system, which is responsible for providing information to work on data governance and description, and modeling of raw data. It includes a repository where all relevant annotations can be machine readable by using an RDF ontology (Resource Description Framework (https://www.w3.org/RDF/ (accessed on 1 February 2026)). This ontology contains characteristics of the input data, such as their attributes and sources. Differently,Solid [14] aims to integrate heterogeneous data under the same data model. Using RDF in conjunction with OWL (Web Ontology

Language (https://www.w3.org/OWL/ (accessed on 1 February 2026)), semantics can be associated with individual schemes facilitating their integration. The data layer of Solids can be seen as storing large batches of RDF data, where triplets are manipulated using a binary representation of RDF. The index layer, which sits on top of the data layer, provides efficient queries to these batches using typically SPARQL (https://www.w3.org/TR/rdf-sparql-query/ (accessed on 1 February 2026)). Finally, Polystore [15] unifies queries when there are multiple heterogeneous storage engines with different data and query models. It organizes data in different islands, where each one represents a category of storage engines that provide a single data model and appropriate query languages to manipulate data found in that island. For example, a relational island can be a collection of traditional database management systems, such as MySQL (https://www.mysql.com/ (accessed on 1 February 2026)) or Postgres (https://www.postgresql.org/ (accessed on 1 February 2026)). It is also possible for an engine to appear in multiple categories and therefore on multiple islands. The idea is that a user can consult an island through its corresponding query language.

Close to domain cases, the contribution in [5] shows how the lack of reuse of software artifacts across different projects can be overcome by selecting an application case that shares considerable commonalities across different projects and providing a project workflow using the example of anomaly detection for process plants.

Although all these efforts address reusability and/or semantics during BDSs development, their integration is scares. In this sense, the novelty of our proposal relies on a variety of identification processes by adding contextual semantics that guide the whole development of BDSs. This allows us to handle the concept of reusability from the definition of case study hypotheses to the visualization of the results, and to use support tools that enable the retrieval of similar variants in different application cases.

Our Approach in a Nutshell

Our methodology mirrors some foundations of software product line (SPL) development [16] (Figure 1). In particular, we apply a two-phase engineering approach to create commonalities and variabilities within a domain: (1) the domain phase, which aims to identify, model, and implement common services of the domain; and (2) the application phase, which focuses on selecting the common services defined in the previous phase and implementing the specific services of the application being developed. Additionally, we define the concept of variety as the main mechanism for flexibility and reuse, but it is restricted to four categories, which are aligned with the main activities of our methodology, allowing each to define common and variable services according to the type of requirement.

The categories are defined as follows: (1) source variety, which detects and defines different data structures, acquisition techniques, etc., during the collection activity; (2) content variety, which focuses on the definition of relevant variables for the business goals to be achieved; mostly considering available data sources during preparation, which focuses on selecting features as well as on traditional data transformation (process variety); (3) process variety, which defines methods and techniques needed to transform data during preparation, and detects variations in data analysis techniques during the analytics activity; and (4) context variety, which allow developers to identify domain variations that may constrain or affect the results of the analysis during the development of the whole BDS [17]. This last variety constitutes the basis of the methodology, making it possible to define the characteristics of the contexts that influence the other varieties. For example, within the agriculture domain, a context variety can be based on regions or seasons, as the systems to be built are dependent on this context to generate different results by applying similar process varieties.

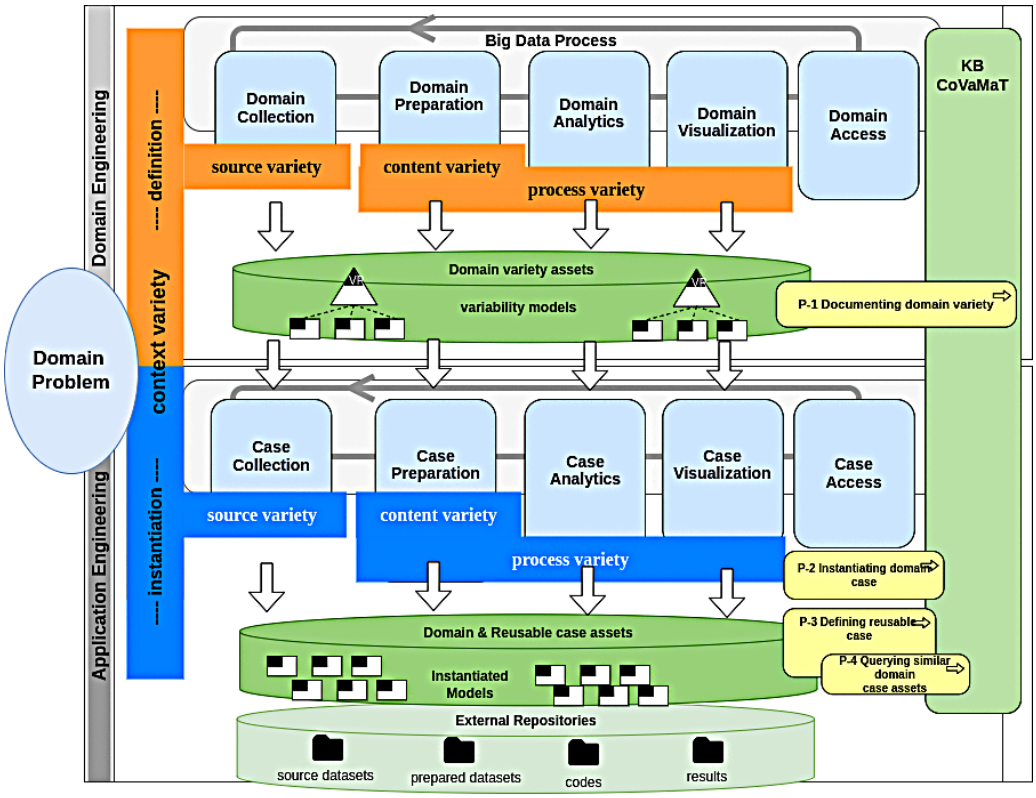


Figure 1. Main components of our reuse-oriented big data methodology.

Models are stored in a knowledge base, named Context-based Variety Management Tool (CoVaMaT) [17]. The domain varieties are stored as domain variety assets by the P-1 process (Documenting domain variety) of CoVaMaT. These models use a simplified version of the Orthogonal Variability Model (OVM) notation [16].

During the application engineering phase, the activities include the same tasks as before but for the development of a specific case. In this phase, firstly domain case assets are generated by instantiating variation points of the domain (defined during domain engineering) and storing these cases ready for reuse. They are created by the P-2 process (Instantiating domain case) of CoVaMaT. Once domain cases exist, reuse case assets are generated by searching for similarity and reusing those stored domain case assets. These activities are done by the P-3 (Defining reusable case) and P-4 (Querying similar domain or case assets) processes. At the same time, during the application engineering phase we also associate the instantiated variants to the files used/generated during the execution of the previous activities of the big data process (datasets, files, codes, etc.).

3. Materials and Methods

In this section, we briefly describe the domain and our previous development of cases, along with the supporting resources already created.

3.1. The Domain and the Analysis of Groundwater Recharge

Groundwater is an underground geological formation that stores fresh water, serving as a vital source for humans and a variety of ecosystems. The process by which groundwater is replenished or recharged is a critical aspect of the global water cycle, involving several stages [18].

To understand how groundwater is recharged, it is necessary to first understand the concept. This term refers to the natural process by which surface water moves into the subsurface. Essentially, there are two main ways in which this process occurs: (1) direct

recharge—this happens when rainwater or runoff infiltrates the soil and reaches the groundwater; and (2) indirect recharge—this is the process by which water from rivers, lakes, and streams seep downward and reach the groundwater. In addition to natural processes, humans have also developed methods to aid groundwater recharge, known as artificial groundwater recharge. This process includes techniques such as direct infiltration, where rainwater is channeled to the groundwater's surface, and induced recharge, where surface or treated water is pumped into infiltration wells. The water table is the upper level of an underground surface in which the soil or rocks are permanently saturated with water. The water table fluctuates both with the seasons and from year to year because it is affected by climatic variations and by the amount of precipitation used by vegetation. It also is affected by withdrawing excessive amounts of water from wells or by recharging them artificially [19].

There are several works in the literature about groundwater recharge analysis, focusing mainly on the factors that influence its increase or decrease [20–22]. For example, in [20], the authors used data extracted from piezometers (a piezometer is a geotechnical sensor used to measure pore water pressure (piezometric level) in the ground) and spatial and temporal variables to find the most influential factors on the groundwater level in an unconfined chalky aquifer of Northern France. Next, using an ANN technique, they estimated the groundwater level. In the work presented in [21], the authors compare two techniques for water table forecasting: an adaptive neurofuzzy inference system (ANFIS), and an ANN. They used data extracted from piezometers located in three areas along the Danube River (between the towns of Kovin and Dubovac in Serbia), as well as weather variables, such as temperature, evapotranspiration, and precipitation, and the levels of both the Danube and a channel that crosses the entire study area. The comparison showed that both techniques achieved similar performance.

Additionally, the work presented in [23] compares various models of machine learning (ML) and deep learning (SVM, generalized regression neural network, decision tree—DT, random forest—RF, convolutional neural network, etc.) to predict the spatial and temporal dynamics of groundwater in the Tarim River area (China), which is an extremely arid region. The inputs were data extracted from 74 piezometers considering the distance of each of them to the river channel, and various weather variables. The results show that Random Forest generated the best results for estimating groundwater levels.

A more recent study compares ML models but using different inputs [24]. In this case, the authors propose using the GRACE (Gravity Recovery and Climate Experiment) satellite to collect data on groundwater levels in the South Khorasan Province, Iran. Specifically, they applied three ML techniques (SVM, RF and DT) to predict groundwater level fluctuations in an arid region. The results show the best performance for DT models.

Next, ref. [25] evaluated the vulnerability of groundwater to identify areas most susceptible to contamination in Beijing, China. The input data included information from 288 groundwater monitoring stations, precipitation records, five topographic classes, land use types, river network density (flow direction and accumulation), groundwater extraction and land management changes. All these variables were assessed using variations in the DRASTIC framework [26].

Finally, two systematic reviews analyzed both neural networks and ML approaches for groundwater level prediction [27,28]. These reviews extract useful information from more than 200 primary studies on aspects such as area of study, input variables, techniques' hyperparameters, performance metrics, software programs, etc.

3.2. Reusable Domain Assets Already Available: Factors Influencing the Groundwater Level in Different Irrigation Periods

In a previous work [7], we described two application cases developed jointly by expert users of the National Institute of Agricultural Technology (INTA) at the experimental station in Alto Valle of Río Negro and Neuquén, Argentina (https://www.argentina.gob.ar/inta/cr-patagonia-norte/eea-altovalle (accessed on 1 February 2026)) and members of our research group. Alto Valle is an extensive region located in the northern part of Patagonia, which includes Neuquén, Limay and Río Negro Rivers. It is a productive area dedicated to the cultivation of pears and apples, with most of the production destined for export and the concentrated juice industry. In particular, in that work, we defined two main hypotheses related to the *factors that could influence the behavior of groundwater, emphasizing weather variables, river flow, distances between the river and groundwater, and irrigation*. Our report aimed to generate variable domain assets and reuse them, subsequently detecting variations due to context. Our objective was not to compare data analytics techniques, but rather to initiate asset loading and analyze their potential for reuse in different contexts.

Therefore, during the development of these two cases, we created several reusable assets following the application of the proposed methodology (Figure 1). During the domain phase, we generated the precision agriculture domain asset together with the variants defined for the four varieties (source, content, process, and context). At this point, the varieties were represented abstractly by the variability models.

Next, during the engineering phase, we instantiated the variants defined for the domain assets to create two domain cases based on the context variety by representing two periods (no irrigation and irrigation), corresponding to fall–winter and spring–summer, respectively. In Figure 2, we show part of the assets stored in CoVaMaT. On the one hand, the DC influences non-irrigation VR asset is associated with different instantiated variants of the domain case (DC). For example, the weather station Villa Regina represents one of the variants instantiated during the collection of the source variety category, which in turn is associated with the weather_station_VillaRegina.csv file. On the other hand, the RC influences irrigation VR asset instantiates variants similarly in the reuse case (RC), but it is associated with a different set of files, such as those corresponding to the irrigation months and the results specifically obtained in this case.

The main findings from verifying the hypotheses of these application cases indicated that, during the non-irrigation period, both weather variables (such as humidity, dew point, and rainfall) and river flow influenced the fluctuations in groundwater levels. Specifically, piezometers located closer to the river showed a greater influence, while those farther away exhibited a decrease in these effects. However, in the latter cases, as the influence of river flow decreased, the impact of weather variables became more significant, increasing their contribution to groundwater fluctuations.

On the other hand, for the irrigation period, the conclusion was completely different. Irrigation generated a new situation in which water itself became the only variable influencing the groundwater level, while the contributions of weather and river flow variables were very low. It is important to highlight that irrigation in this region uses a system known as flood irrigation, or surface flooding. It works by opening canals that allow water to flow by gravity and spread across the orchard or field, saturating the soil. In this way, the study showed that this flood irrigation overshadowed the influence of the other variables. Therefore, by reusing previous knowledge and stored assets, we could explore domain problems more deeply and obtained a better understanding of the behavior of the studied variables.

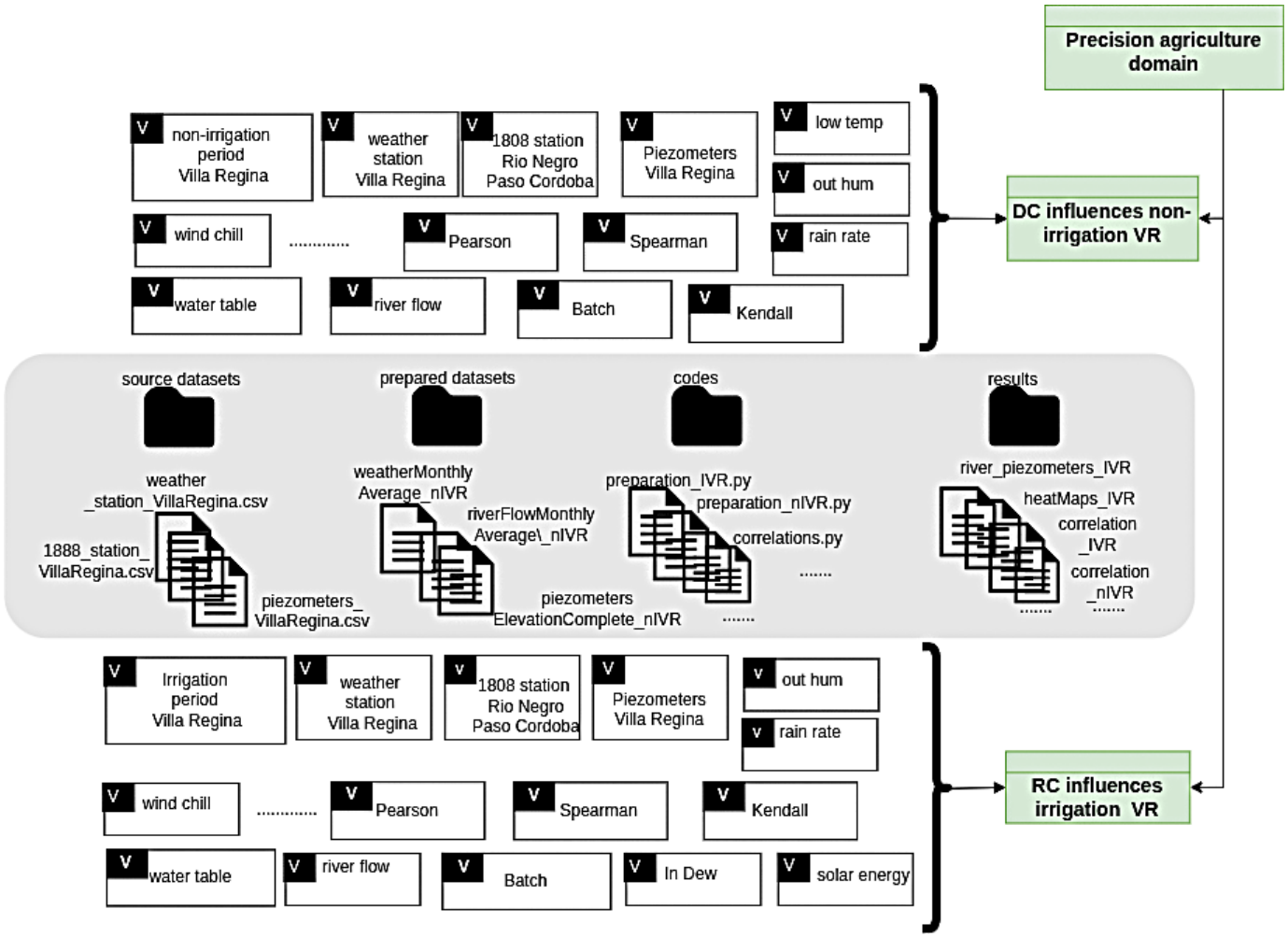


Figure 2. Part of the assets stored in CoVaMaT for the two previous application cases.

4. Detecting Water Table Level Influences Through Reusing Domain Assets

In order to analyze the feasibility of our approach, and given the assets already stored, this section introduces two reuse cases for predicting fluctuations in groundwater levels. The reuse process will be introduced abstractly through the use of BPMN (https://www.bpmn.org/ (accessed on 1 February 2026)) diagrams, which will be exemplified by the two cases and three hypotheses.

Firstly, Figure 3 shows an abstract view of the entire process. The work was carried out in close collaboration with domain experts, who, under certain conditions, generated a data analysis requirement. This requirement is addressed through process P3 (Develop a Reuse Case), which begins by creating the case. This involves formulating hypotheses and retrieving domain assets from similar cases. The process then continues by instantiating the case; that is, performing the analysis while associating the identified varieties (source, content, process) or new varieties specific to the case. Finally, the results are evaluated to determine whether to continue the process or if they are sufficiently satisfactory (a BPMN's complex gateway is used before the process ends).

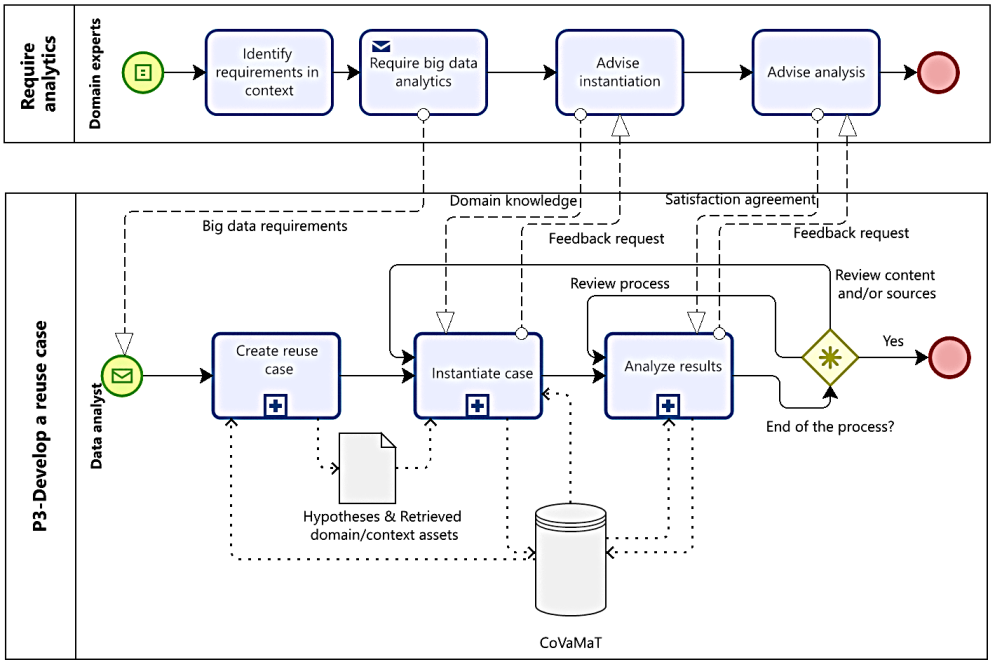


Figure 3. Overview of the entire P3 (development a reuse case) process.

Let us examine each of these activities in our domain problem.

4.1. Create Reuse Case

The diagram in Figure 4 breaks down the activity into steps to create the reuse case. Once the domain and context are identified—in our case, agrometeorology in the context surrounding the Alto Valle INTA experimental station—the objectives and hypotheses of the case are defined.

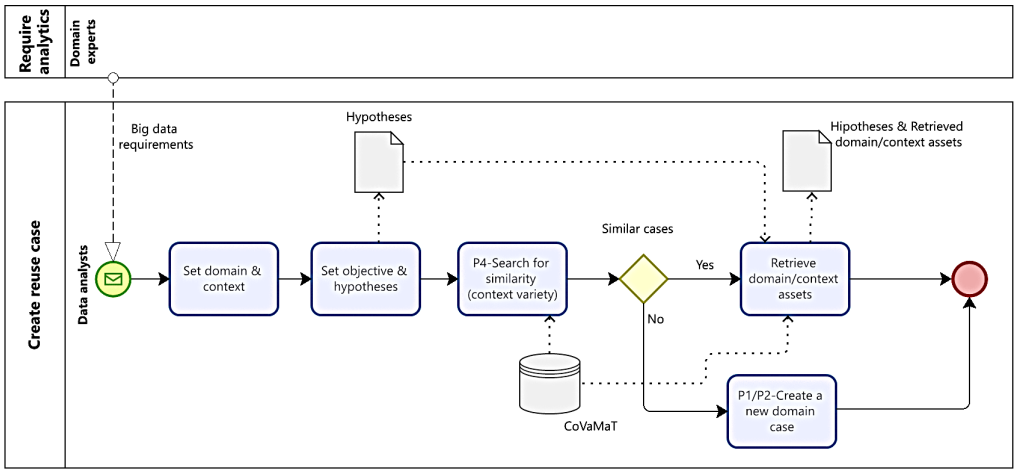


Figure 4. Create reuse case.

Based on knowledge from domain experts and a literature review [7], we defined the following objective:

Identifying the water table recharge influences and predicting fluctuations under different conditions.

As shown in Figure 4, after determining the domain objective, the process continues by defining the hypotheses and retrieving the identification and context of similar cases (Retrieve domain/context assets). Finally, if; for example, no similar context exists, the

system allows the creation of varieties using processes P1 and P2 (Create new domain case). Let us look at these activities in the two cases under study.

4.1.1. Creating the Cases: nIVR and IVR

To identify contextual influences, this objective was addressed in two different periods: the irrigated period (IVR) and the non-irrigated period (nIVR). As mentioned in Section 3.2, CoVaMaT stores the varieties from two cases developed during these two periods for a single location (VR-Villa Regina) [7]. In these cases, the analysis was limited to searching for correlations between meteorological variables, the flow rate of the Río Negro (which crosses the area distributing water through irrigation canals), and variations in water table level. The cases were created from scratch using processes P1 (Document Variety) and P2 (Instantiate Case); and now it is time to reuse the existing data and determine if the stored assets can be reused and useful for analyzing the following hypotheses, which focus on predicting variations in water table level.

4.1.2. Common Hypotheses for Both Contextual Cases (nIVR and IVR)

Hypothesis 1. *Is it possible to efficiently predict the water table level of each piezometer using the flow rate variable of the Río Negro River?*

Hypothesis 2. *Is it also possible to efficiently predict the water table level of each piezometer using weather variables?*

Hypothesis 3. *Is it also possible to efficiently predict the water table level of each piezometer using the most influential weather variables discovered in previous studies?*

4.1.3. An Additional Hypothesis for the Second, Case (IVR)

In this second case (the case of the non-irrigation period was carried out first, so we will refer to it in the article as the first case or nIVR; consistently, the case of the irrigation period will be referred to as the second case or IVR), we incorporated a new requirement provided by domain experts. This requirement aimed to assess whether models could efficiently predict water table levels with the same weather and river flow variables as the period of non-irrigation, but now during the irrigation period (Set objective and hypotheses activity in Figure 4). Thus, we defined a new hypothesis:

Hypothesis 4. *Are predictive models developed for the non-irrigation period also effective during the irrigation period?*

Before starting with processing the hypotheses, we looked for resources stored in CoVaMaT (P4-Search for similarity activity in Figure 4). At the time of creating this new case, the recently created nIVR reuse case had already been completed. Therefore, CoVaMaT stored data from the two previous cases (Section 3.2) and nIVR (with all its varieties already stored).

4.2. Instantiate the Case

Figure 5 shows the instantiation of a case through several sequential states (from data collection to analytics) in a simplified version of reality to illustrate the identification and reuse of different types of varieties.

First, an identification activity stores the reuse case in CoVaMaT (note that for readability, CoVaMaT appears twice in the figure below), preparing the ground for the respective instantiations. Let us examine each of these in detail for the cases.

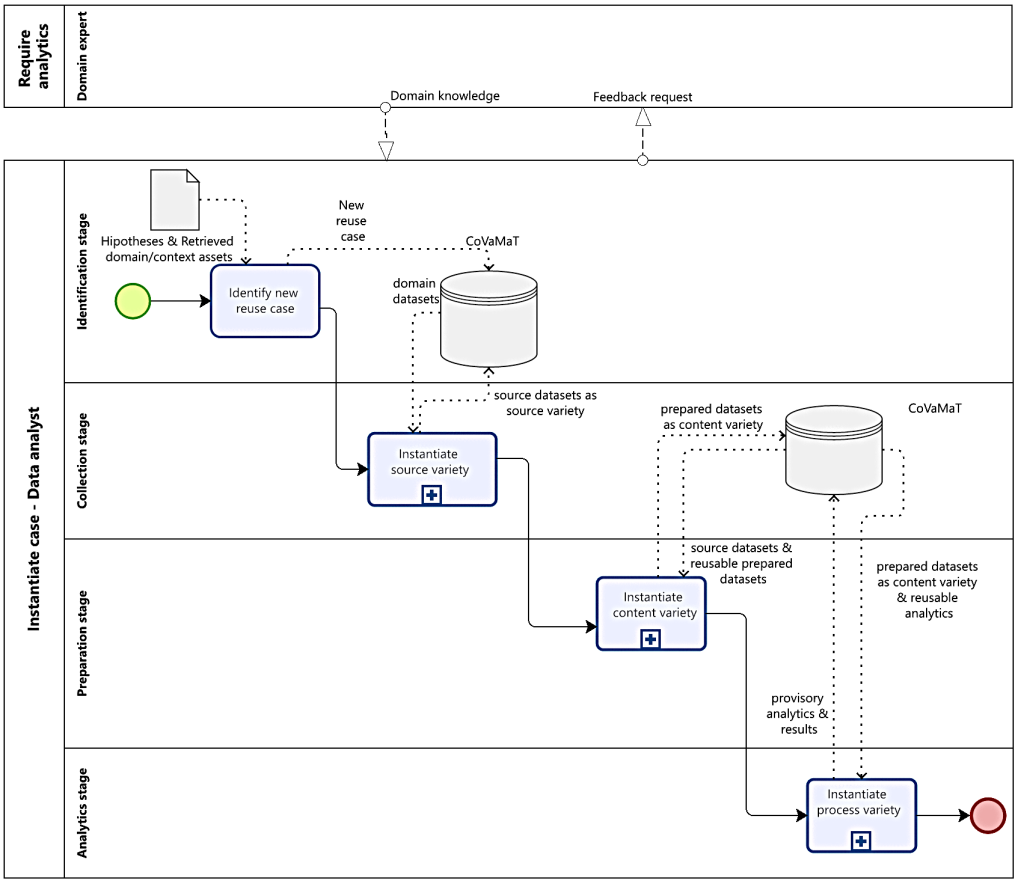


Figure 5. Instantiate reuse case.

4.2.1. Instantiate Source Variety

As shown in Figure 6, relevant existing datasets are identified during the collection stage. This activity may require assistance from domain experts to add, extend, or interpret information from the perspective of the new hypotheses. As in all situations, if a relevant dataset does not already exist, it can be created from scratch using processes P1/P2 (P1/P2 create a new source variety (raw and clean datasets) activity in Figure 6 (Note that this activity potentially takes place in two different stages: collection and preparation, a fact omitted from the diagram for simplicity)). Once all datasets have been collected in both their raw and cleaned versions (after a process that; for example, performs some treatment on null data), they are associated with the reuse case.

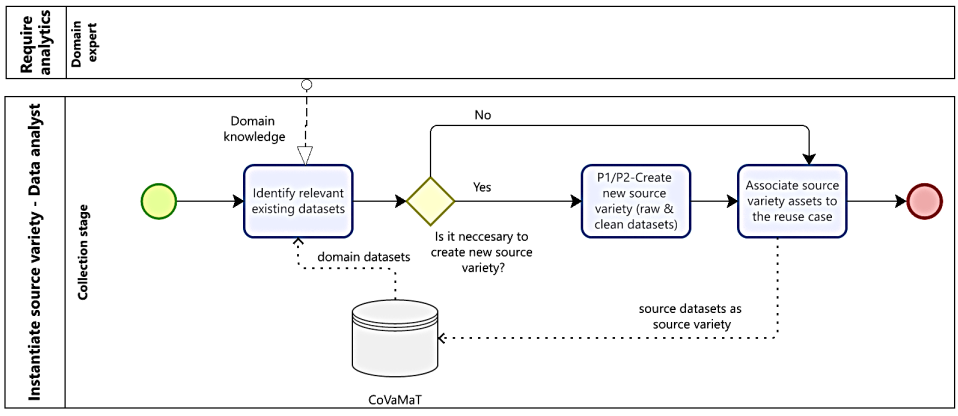


Figure 6. Instantiate source variety.

First Case: The Non-Irrigation Period in Villa Regina (nIVR)

The activity of identifying relevant existing datasets allowed us the retrieval of the following domain datasets from CoVaMaT:

- 1888_station_VillaRegina.csv: Contains a daily report of the flow variables of the Río Negro River. Specifically, it contains four variables: date, height, IGN (IGN refers to the National Geographic Institute, which provides the official elevation reference system) elevation, and corrected IGN elevation.
- weather_station_VillaRegina.csv: Contains 33 variables measured at 10 min intervals. For example, some of the climate variables are out humidity, atmospheric pressure, solar radiation, rain rate, wind speed, low temperature, etc.
- piezometers_VillaRegina.csv: Contain a monthly report of each of the 77 piezometers located in Villa Regina region. These data are semi-structured and contain four variables: a piezometer identifier, elevation, month, and groundwater level in *masl* (meters above sea level).

All datasets included data from January 2010 to December 2023.

At the same time, we analyzed the prepared datasets and notebooks (python notebooks codes) applied to the preparation activity and already stored in CoVaMaT. For example, we observed that the piezometer dataset had been cleaned and that the null values and outliers had been imputed. In addition, only 15 piezometres were selected because, according to expert users, the influence of the river was limited to a radius of 2 km from the flow. The name of this prepared dataset is piezometersElevationComplete_nIVR.

Next, regarding to the weather dataset, the documentation in CoVaMaT showed that a set of 14 weather variables was cleaned, formatted, and normalized to include a monthly average. The name of this prepared dataset is weatherMonthlyAverage_nIVR.

Finally, we also retrieved the riverFlowMonthlyAverage_nIVR dataset, which includes the flow rate variables of the Río Negro River normalized to the same monthly average.

It is important to highlight that all these prepared datasets contain records from May to August of each year, which is precisely the period without irrigation.

All these datasets were associated with the reuse case nIVR (Associate source variety assets to the reuse case in Figure 6).

Second Case: The Irrigation Period in Villa Regina (IVR)

During the *collection stage* we used the same source and prepared datasets stored for the RC influences irrigation VR asset (previous work, Section 3.2). They are the same as in the non-irrigation case, but cover a different period: from September to April, 2010–2023.

The names of the retrieved prepared datasets are as follows: piezometersElevation-Complete_IVR, weatherMonthlyAverage_nIVR, andriverFlowMonthlyAverage_IVR; all of them are normalized to the same monthly average.

4.3. Instantiate Content Variety

During the preparation stage, Figure 7 illustrates the process of instantiating the content variety, which begins by identifying relevant variables according to the case hypotheses. This can involve feature analysis, creating value from the initial data (e.g., suggestions for generating new variables from existing data or by applying domain knowledge to improve a model's predictive capacity), correlation analysis to detect the strongest influences, and so on.

Starting with the source data (reused or not) in the datasets, an identification process is initiated based on knowledge from previous cases and assistance from domain experts, enabling the generation of prepared content. Contextual knowledge stored through previous

domain cases in CoVaMaT (such as conclusions drawn from past cases) can have a significant influence on this preparation. As the figure below shows, the prepared datasets are associated with the reuse case as a content variety.

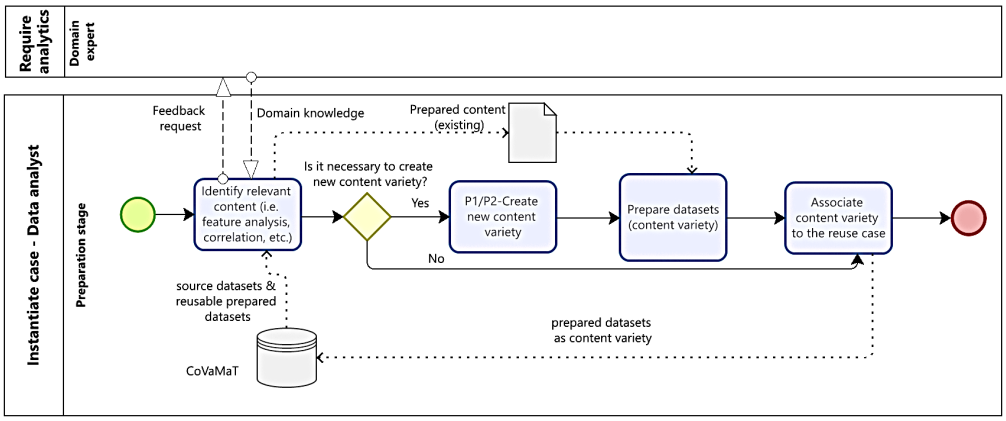


Figure 7. Instantiate content variety.

4.3.1. First Case: The Non-Irrigation Period in Villa Regina (nIVR)

Considering Hypothesis 1, (Is it possible to efficiently predict the water table level of each piezometer using the flow rate variable of the Río Negro River?) we used the results stored in CoVaMaT about the influence of the river on groundwater levels. Specifically, we used the riverFlowMonthlyAverage_nIVR and piezometersElevationComplete_nIVR datasets.

In CoVaMaT, previous cases identified some of the following conclusions for the non-irrigation period:

- The most influential weather variables associated with fluctuations in water table levels were dew point, humidity, and rain rate. These influences were stronger in the piezometers located farther from the course of the Río Negro River.
- The river flow also shown strong influences, especially in piezometers located closest to the course of the Río Negro River.
- Models based on dew point, humidity, and rainfall were highly effective in forecasting water table levels.
- Models based on river flow were also very effective in forecasting water table levels.
- Models based on weather variables shown slightly better performance than those based on river flow.

4.3.2. Second Case: The Irrigation Period in Villa Regina (IVR)

In this case, we should consider that during the irrigation period, the following conclusions were identified:

- Weather variables and river flow only shown very low influence on water table fluctuations.
- The previous results showed that flood irrigation minimized or eliminated the influence of any other analyzed variables.

4.4. Instantiate Process Variety

Instantiating the process variety involves iteratively performing calculations that adjust parameters during the execution of the analytics, generating provisional results. In practice, these results can be evaluated with various quality measures immediately after calculation or later, after obtaining a set of results. Figure 3 shows the case where the results analysis is performed later; however, note that this is a simplification of the execution sequence that may occur in practice.

During the analytics stage, Figure 8 shows the instantiation of the provisory results for process variety, beginning with the identification of two complementary elements: the previously prepared datasets and analytics that can be reused from previously identified cases. It may occur that the previous analytics are insufficient and it becomes necessary to add new analysis methods; in that case, processes P1/P2 will create a new process variety. The application of analytics and subsequent visualization of results will generate iterative variations that will produce a set of provisional results, awaiting evaluation through quality measures.

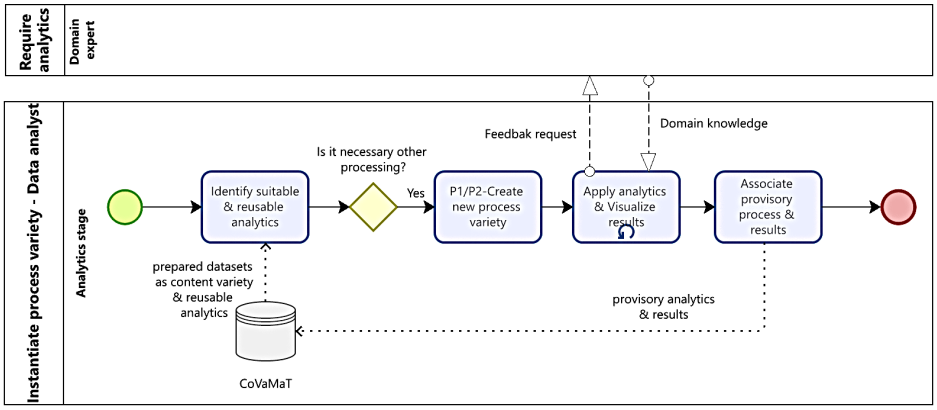


Figure 8. Instantiate process variety.

4.4.1. First Case: The Non-Irrigation Period in Villa Regina (nIVR)

To address Hypothesis 1 (Is it possible to efficiently predict the water table level of each piezometer using the flow rate variable of the Río Negro River?) and based on the information of the most affected piezometers, we applied three ML techniques to predict groundwater levels. For the selection of techniques, we based our decision on the systematic review presented in [28]. According to this article, which analyzed 223 primary studies, the ANN, SVM and LSTM techniques were the most used and applied. Therefore, the hypothesis was evaluated according to those ML techniques with specific configurations.

To build the ML models, we used only the flow rate of the Río Negro River as the input variable (independent variable), along with the piezometer identifiers as categorical variables, allowing the model to distinguish between the different measurement points. The output variable (dependent variable) was the water table level. We also organized the observations into 4-month time windows, using the months of May, June, July, and August of each year. With this structure, the model predicted the groundwater level based on the previous 4 months. The data were split into 80% for training and 20% for testing, leaving the 2023 data for validation with actual values that were not used during training. At the same time, for each ML technique, we analyzed and validated several configuration. We used TensorFlow (https://www.tensorflow.org/?hl=es-419 (accessed on 1 February 2026)) and Keras (https://keras.io/ (accessed on 1 February 2026)) for the implementation.

Firstly, we trained a recurrent neural network (RNN) using an LSTM, specifically three different configurations. In particular, in Table 1, we present the hyperparameters of each configuration. In the first configuration, a single LSTM layer with 32 units was adopted to provide a simple, low-complexity architecture suitable for the limited dataset size. The layer used the tanh activation function and a dropout rate of 20% to reduce the risk of overfitting. Model optimization was performed using the Adam optimizer with a learning rate of 0.0005 to improve training stability. Early stopping was applied, configured to halt training if the validation loss did not improve for 15 consecutive epochs, with the best-performing weights restored. Training was carried out using a batch size of 16 and a maximum of 200 epochs.

Table 1. Hyperparameters used on each of the three configurations of the LSTM model.

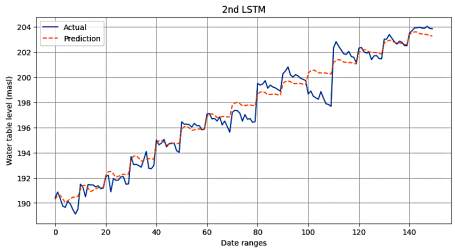
Hyperparameters	1st Configuration	2nd Configuration	3rd Configuration
Number of LSTM layers	1	2	3
Number of units per layer	32	128–64	256–128–64
Batch size	16	16	32
Number of epochs	200	200	200
Learning rate	0.0005	0.001	0.001
Activation function (LSTM)	tanh	tanh	tanh
Dropout rate (LSTM)	0.2	0.3	0.3
Optimization	Adam	Adam	Adam

For the second configuration, we changed to a two-layer LSTM architecture, where the first layer had 128 units and returned the complete output sequence to the next LSTM layer, which had 64 units. We changed the learning rate to the common default value of 0.001, and the rest of the parameters remained unchanged. Finally, the third configuration added an additional LSTM layer with 256, 128, and 64 units, respectively, with the aim of assessing whether increased model depth could improve the representation of temporal dependencies.

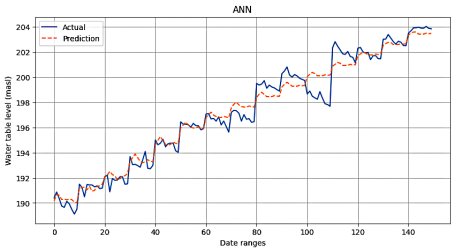
Next, we trained a feed-forward ANN using the MLPRegressor algorithm. The best configuration of hyperparameters was one hidden layer with 20 neurons, the relu activation function, maximum training iterations of 500, and thelbfgs optimizer with a learning rate of 0.01.

Finally, we implemented a SVM regression model using the SVR algorithm. Here, the best hyperparameters were a rbf kernel, a penalty of 100, an epsilon of 0.2 and a gamma of 0.01.

As an illustrative analysis, Figure 9 shows a comparison between the observed values (blue) and the predicted values (orange) obtained from the three best ML models considered: the second LSTM configuration (selected as the best-performing LSTM), the ANN, and the SVM. In particular, the best results were obtained by the ANN technique using MLPRegressor algorithm, as discussed in Section 5.



(**a**) 2nd LSTM configuration



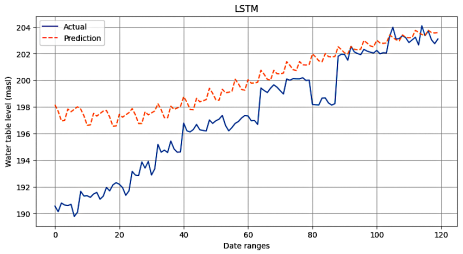
(**b**) ANN configuration

(**c**) SVM configuration

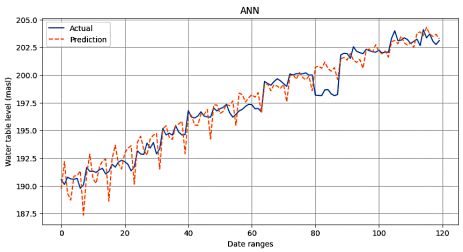
Figure 9. Comparison of the actual values (blue) and predicted values (orange) of the three ML models for river flow rate and water table levels.

Considering Hypothesis 2, (Is it also possible to efficiently predict the water table level of each piezometer using weather variables?) we used weatherMonthlyAverage_nIVR and piezometersElevationComplete_nIVR datasets. The selection of weather variables was guided by expert knowledge and the systematic review reported in [27]. The review identifies evaporation, temperature, precipitation, rainfall, evapotranspiration, humidity, and wind speed as the most frequently used predictors in groundwater level modeling. Furthermore, domain experts recommended excluding weather variables such as heat index and wind chill. In this way, the variables selected as inputs in this hypothesis were: Hi Temp, Low Temp, Out Hum, Dew Pt., Wind Speed, Bar, Rain Rate, Solar Rad., and ET. The output (dependent) variable was again the water table level. To build the different models we tested the same configurations as for Hypothesis 1.

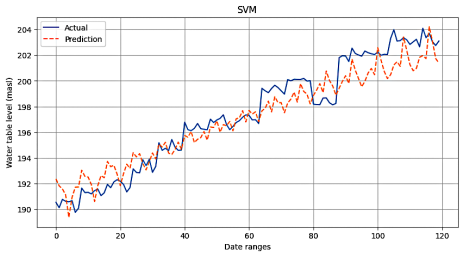
Figure 10 shows graphically the comparison between the observed values (blue) and the predicted values (orange) obtained from the three best ML models considered.



(**a**) LSTM configuration



(**b**) ANN configuration



(**c**) SVM configuration

Figure 10. Comparison of the actual values (in blue), and predicted values (in orange) of the three ML models for weather variables and water table levels.

For Hypothesis 3 (Is it also possible to efficiently predict the water table level of each piezometer using the most influential weather variables?), we used the results stored in CoVaMaT regarding the influence of weather variables on groundwater levels. Previous cases identified a moderate relationship between humidity, dew point and rain rate with respect to water table levels; therefore, we used these weather variables as input features for the LSTM model. The target variable was again the water table level. In this case, we applied the same hyperparameters as for Hypothesis 2.

Finally, we designed an interactive application to allow users to test the trained LSTM models. As shown in Figure 11 (the form is in Spanish because it is used by expert users of INTA), the form allows users to enter specific values for the weather variables of humidity, dew point and rain rate as a basis for the prediction.

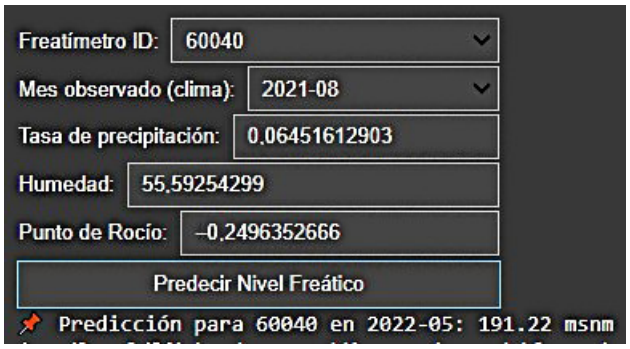


Figure 11. Interactive form with specific input data for the prediction of the groundwater level.

At the same time, users must enter the piezometer identification and the month and year for which the prediction should be performed. In the figure, the prediction was made for the piezometer 60040 for August 2021. In this case, based on the rain rate (0.0645), humidity (55.5925), dew point (−0.2496), the model predicted a value of 191.22 *masl*. When comparing this value with the actual one, the prediction obtained a difference in only 0.05 *masl*, as we can see in Table 2.

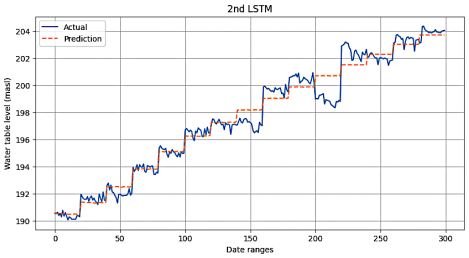
Table 2. Comparison between the predictive value and the actual value.

Piezometer ID	Prediction (*masl*)	Actual Value (*masl*)	Difference (*masl*)
60040	191.22	191.27	0.05

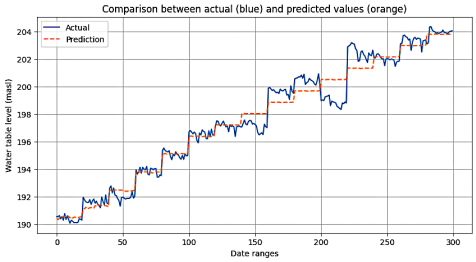
4.4.2. Second Case: The Irrigation Period in Villa Regina (IVR)

For Hypotheses 1, 2, and 3, we reused the best-performing model configurations identified in the previous case, which were the second LSTM configuration (shown in Tables 1 and 3), ANN and SVM.

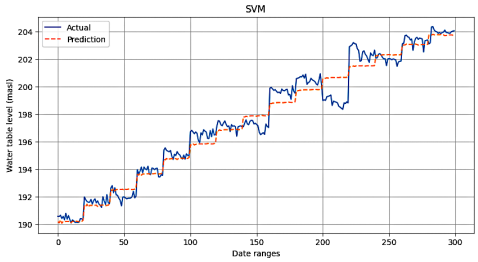
For simplicity, we show graphically here only the results of Hypothesis 1, as shown in Figure 12. In particular, the best results were obtained by the ANN technique using MLPRegressor algorithm, which will be discussed in the next section.



(**a**) 2nd LSTM configuration



(**b**) ANN configuration



(**c**) SVM configuration

Figure 12. Comparison of the actual values (in blue), and predicted values (in orange) of the three ML models for river flow and water table levels.

Table 3. Metrics applied to the ML models for the flow rate of the river and water levels for $H1$.

					Improvements in Metrics			
Model	**MSE**	**MAE**	**R^2**	**RMSE**	**MSE (%)**	**MAE (%)**	**ΔR^2**	**RMSE (%)**
1st LSTM	0.0385	0.1422	0.9636	0.1963	12.47	1.90	0.0046	6.48
2nd LSTM	0.0344	0.1419	0.9675	0.1856	2.03	1.69	0.0007	1.08
3rd LSTM	0.1045	0.2250	0.9013	0.3233	67.75	38.00	0.0669	43.21
ANN	**0.0337**	**0.1395**	**0.9682**	**0.1836**	0.00	0.00	0.0000	0.00
SVM	0.0349	0.1447	0.9670	0.1869	3.44	3.59	0.0012	1.76

5. Analyze Results

To conclude the process (activity analyze results in Figure 13), the results are evaluated using various metrics that, if satisfactory, would allow for the storage of definitive results for the process variety. These metrics and additional information from the case study then form the basis for identifying the next course of action: terminating the entire process or restarting one of its stages, as shown in Figure 3.

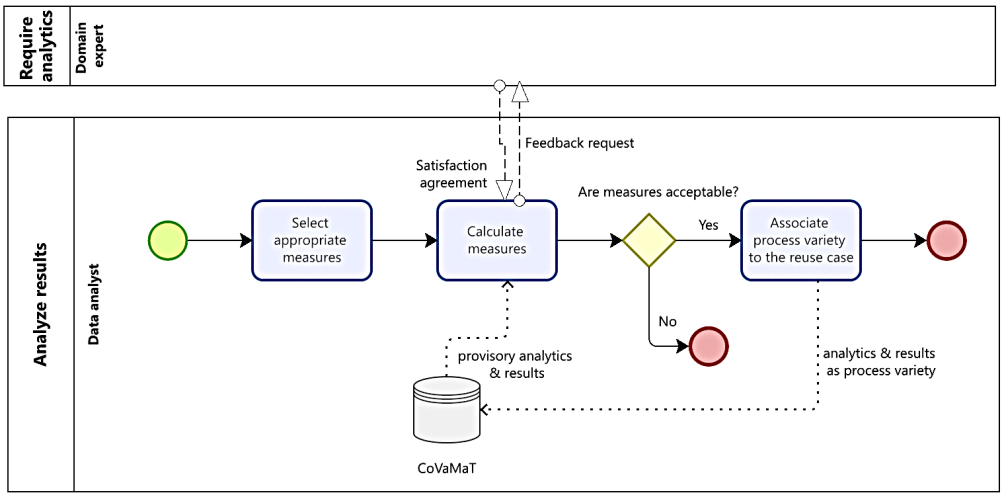


Figure 13. Analyze results.

To validate the models, we applied four performance metrics [27]:

- Mean Absolute Error (MAE): Represents the average absolute difference between the actual and predicted values. Its values range from 0 to $+\infty$, where lower MAE values indicate better model performance. The formula is as follows:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{1}$$

where n is the number of observations, y_i represents the actual value, and $\hat{y}_i$ represents the predicted value, both for the i-th observation. The unit is the same as the target variable (water table level). In this case, *naml*.

- Mean Squared Error (MSE): Represents the average of the squared difference between the original and predicted values, and therefore penalizes large errors more severely. Its values range from 0 to $+\infty$. As with the previous metrics, values closer to 0 indicate better model performance. The formula is as follows:

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

where n, y_i, and $\hat{y}_i$ represents the same as MAE. The unit is the square of the target variable's unit ($masl^2$).

- Root Mean Squared Error (RMSE): Represents the square root of MSE. Its values range from 0 to $+\infty$. The formula is as follows:

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

where n, y_i, and $\hat{y}_i$ represents the same as MAE. Returns the error to the original units of the target variable ($naml$).

- Coefficient of determination (R^2): Refers to how well the model's predictions approximate the true values. It represents the proportion of the variance in the dependent variable (water table level) that is explained by the independent variable (flow rate of the Río Negro River). A value of 1 indicates a perfect fit. The formula is as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4}$$

where $\bar{y}$ is the mean of the observed values.

5.1. First Case: The Non-Irrigation Period in Villa Regina (nIVR)

The analysis to test Hypothesis 1 (Is it possible to efficiently predict the water table level of each piezometer using the flow rate variable of the Río Negro River?) resulted in the values shown in Table 3 for each of the four metrics. To compare the ML models, we computed the relative error reduction with respect to the best-performing model. As MSE, MAE and RMSE are error metrics for which lower values indicate better performance, the percentage improvement of a model i with respect to the best model is defined as (for MSE, for instance):

$$\text{Improvement}_i(\%) = \frac{\text{MSE}_i - \text{MSE}_{\text{best}}}{\text{MSE}_i} \times 100, \tag{5}$$

where MSE_{best} denotes the MSE of the model achieving the lowest error. For example, we can see in Table 3 that ANN obtained the lowest MSE (0.0337). In contrast, the 3rd LSTM presented a higher MSE, resulting in a large relative improvement of 67.75%.

For the case of R^2, model performances are compared using the absolute difference (ΔR^2), defined as the difference between the R^2 value of the best-performing model and any other model. For example, the value $\Delta R^2 = 0.0669$ reported for the 3rd LSTM (Table 3) corresponds to the difference between the ANN ($R^2 = 0.9682$) and the 3rd LSTM ($R^2 = 0.9013$). This indicates that ANN explains an additional 0.0669 of the total variance in water table levels compared to the 3rd LSTM.

The analysis to test Hypothesis 2 (Is it also possible to efficiently predict the water table level of each piezometer using weather variables?) resulted in values shown in Table 4 for each of the four metrics.

Table 4. Metrics and relative improvements for weather variables and water table levels for $H2$.

					Improvements in Metrics			
Model	**MSE**	**MAE**	**R^2**	**RMSE**	**MSE (%)**	**MAE (%)**	**ΔR^2**	**RMSE (%)**
LSTM	0.6745	0.6374	0.8213	0.3404	89.55	68.00	0.1098	22.00
ANN	**0.0705**	**0.2040**	**0.9311**	**0.2655**	0.00	0.00	0.0000	0.00
SVM	0.0939	0.2596	0.9082	0.3064	24.93	21.42	0.0229	13.35

The analysis to test *Hypothesis 3* (Is it also possible to efficiently predict the water table level of each piezometer using the most influential weather variables?) resulted in values shown in Table 5.

Table 5. Metrics and relative improvements for specific weather variables and water table levels for *H*3.

					Improvements in Metrics			
Model	**MSE**	**MAE**	**R^2**	**RMSE**	**MSE (%)**	**MAE (%)**	**ΔR^2**	**RMSE (%)**
LSTM	**0.0329**	**0.1313**	**0.9687**	**0.1813**	0.00	0.00	0.0000	0.00
ANN	0.0482	0.1715	0.9545	0.2195	31.74	23.41	0.0142	17.40
SVM	0.6414	0.6174	0.3944	0.8009	94.87	78.73	0.5743	77.36

In this step, we compared the three hypotheses to analyze the predictive models by considering the flow of the Rio Negro River versus the use of different weather variables. Table 6 presents the results of the three model configurations: the first row corresponds to the ANN implementation using river flow as the input variable (*H*1), the second row to the ANN model using multiple weather variables (*H*2), and the third row to the LSTM model using a selected subset of weather variables (*H*3). Although the differences among the evaluation metrics are relatively small, the LSTM implementation using only three weather variables achieved the best performance for predicting water table levels.

Table 6. Metrics applied to the best ML models for the three hypotheses.

	MSE	**MAE**	**R^2**	**RMSE**
ANN for *H*1	0.0337	0.1395	0.9682	0.1836
ANN for *H*2	0.0705	0.2040	0.9311	0.2655
LSTM for *H*3	0.0329	0.1313	0.9687	0.1813

Finally, we decided to finish the application case and store all the generated resources (variants and files) in CoVaMaT. In Figure 14, we can see a summary of the instantiated variants and their associated files (source and prepared datasets, codes, etc.). All these resources are part of the newRC prediction non-irrigation VR included in the precision agriculture domain.

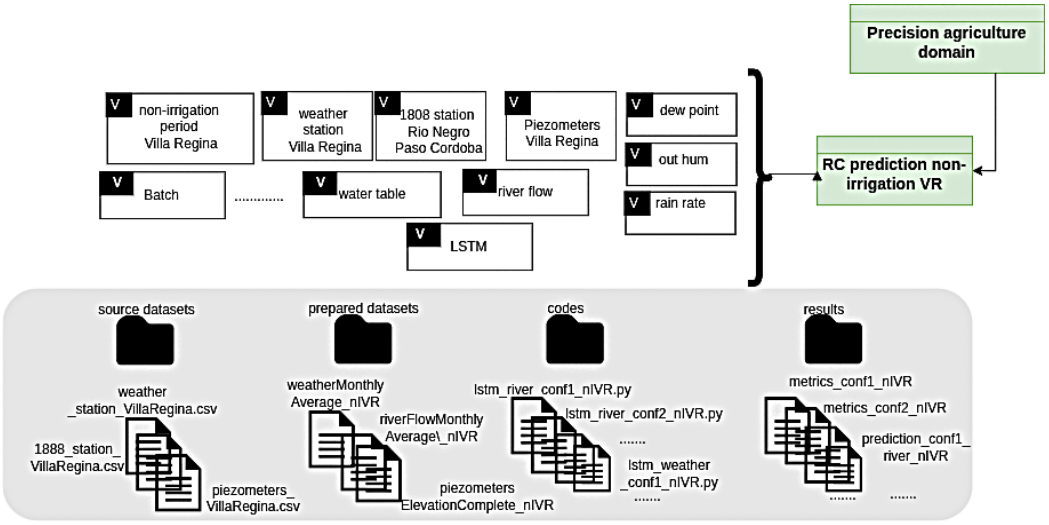


Figure 14. Summary of the resources created for the application case focused on predictions in the non-irrigation period.

5.2. Second Case: The Irrigation Period in Villa Regina (IVR)

Table 7 presents the evaluation metrics obtained for Hypothesis 1, where the first differences with respect to the previous case are observed. Although all models achieved good performance, the second LSTM configuration achieved the best results. In contrast to the previous case, where ANN was the best model, the differences between both models in this scenario are very small.

Table 7. Metrics and relative improvements for flow rate of the river and water table levels during the irrigation period for *H*1.

					Improvements in Metrics			
Model	**MSE**	**MAE**	**R^2**	**RMSE**	**MSE (%)**	**MAE (%)**	**ΔR^2**	**RMSE (%)**
2nd LSTM	**0.0301**	**0.1304**	**0.9702**	**0.1735**	0.00	0.00	0.0000	0.00
ANN	0.0302	0.1314	0.9707	0.1737	0.33	0.76	−0.0005	0.12
SVM	0.0325	0.1411	0.9685	0.1802	7.38	7.59	0.0017	3.72

Next, for the second hypothesis, Table 8 presents the results. Here, the best model performance was again ANN as in the previous case.

Table 8. Metrics and relative improvements for weather variables and water table levels during the irrigation period for *H*2.

					Improvements in Metrics			
Model	**MSE**	**MAE**	**R^2**	**RMSE**	**MSE (%)**	**MAE (%)**	**ΔR^2**	**RMSE (%)**
LSTM	0.0753	0.2201	0.9291	0.3404	41.04	22.94	0.0291	38.07
ANN	**0.0444**	**0.1696**	**0.9582**	**0.2108**	0.00	0.00	0.0000	0.00
SVM	0.0472	0.1723	0.9555	0.2174	5.93	1.57	0.0027	3.04

Finally, the results of the third hypothesis are shown in Table 9. In this case, the best model was ANN, contrary to LSTM for the previous application case.

Table 9. Metrics and relative improvements for specific weather variables and water table levels during the irrigation period for *H*3.

					Improvements in Metrics			
Model	**MSE**	**MAE**	**R^2**	**RMSE**	**MSE (%)**	**MAE (%)**	**ΔR^2**	**RMSE (%)**
LSTM	0.0413	0.1434	0.9611	0.2032	13.80	−6.49	0.0054	7.18
ANN	**0.0356**	**0.1527**	**0.9665**	**0.1886**	0.00	0.00	0.0000	0.00
SVM	0.0373	0.1622	0.9649	0.1932	4.56	5.86	0.0016	2.38

Table 10 presents the results of the three model configurations: the first row corresponds to the second configuration of LSTM implementation using river flow as the input variable (*H*1), the second row to the ANN model using multiple weather variables (*H*2), and the third row to the ANN model using a selected subset of weather variables (*H*3). As in the previous case, the differences among the evaluation metrics were relatively small; however, in this case the second LSTM configuration achieved the best performance for predicting water table levels under *H*1.

Table 10. Metrics applied to the best ML models for the three hypotheses for the irrigation case.

	MSE	**MAE**	**R^2**	**RMSE**
2nd LSTM fo *H*1	0.0301	0.1304	0.9702	0.1735
ANN for *H*2	0.0444	0.1696	0.9582	0.2108
ANN for *H*3	0.0356	0.1527	0.9665	0.1886

Similarly to the previous case, after analyzing the results, we decided to finish the application and store all generated resources (variants and files) in CoVaMaT. Specifically, we documented a new RC prediction irrigation VR asset as part of the precision agriculture domain.

5.3. Comparing the Cases: nIVR Versus IVR

Regarding the last hypothesis of this case, which evaluates whether predictive models developed for the non-irrigation period are also effective during the irrigation period, Table 11 summarizes and compares the results of both periods shown in Tables 6 and 10. As shown, we obtained different techniques as the best according to the context. Specifically, Hypotheses 1 and 3 obtained different ML models that achieved the highest predictive efficiency. As an example, for Hypothesis 1, MSE_{ANN} = 0.0337 for the first case; meanwhile, the best result for the second case was MSE_{LSTM} = 0.0307. Improvements for predictions when selecting different models have already been reported when analyzing the previous hypotheses (Tables 3–9).

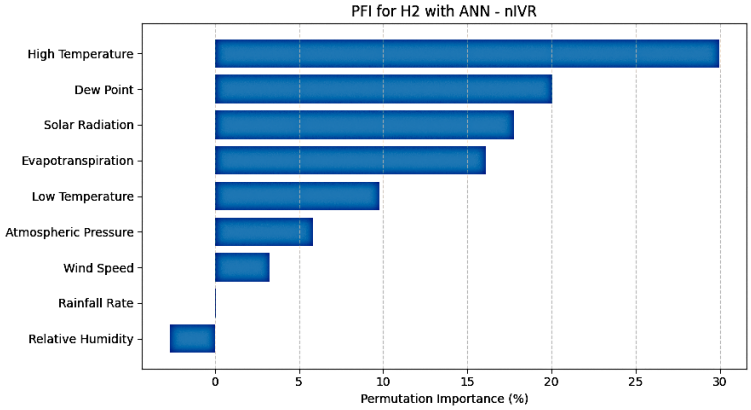
On the other hand, to analyze Hypothesis 4, we evaluated, for each application case and its best-performing ML model, the contribution of individual input variables to the predictions. To do so, we applied the permutation feature importance (PFI) [29], which is a well-known and simple post hoc method for quantifying the contribution of each input variable to the predictions. As it is model-agnostic, it can be applied to any predictive model regardless of its internal architecture.

To perform the evaluation we analyzed Hypotheses 2 and 3 for each case. Regarding the second hypothesis, Figure 15 shows that features' importance of the ANN models are different during the two periods. For instance, during the non-irrigation period, high temperature is the most influential variable, followed by dew point, solar radiation, and other meteorological factors. In contrast, in the IVR case, although high temperature remains the most influential variable, the relative importance of the remaining weather variables changes.

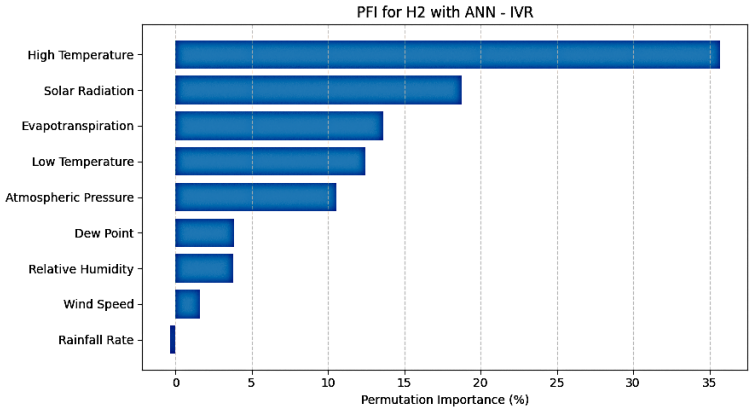
Table 11. Comparison of the best ML models for each hypothesis and application case.

Hypotheses	Application Case nIVR				Application Case IVR			
	MSE	**MAE**	**R^2**	**RMSE**	**MSE**	**MAE**	**R^2**	**RMSE**
*H*1	ANN				LSTM			
	0.0337	0.1395	0.9682	0.1836	0.0301	0.1304	0.9702	0.1735
*H*2	ANN				ANN			
	0.0705	0.2040	0.9311	0.2655	0.0444	0.1696	0.9582	0.2108
*H*3	LSTM				ANN			
	0.0329	0.1313	0.9687	0.1813	0.0356	0.1527	0.9665	0.1886

Finally, we performed the same evaluation for Hypothesis 3 in both cases. In this hypothesis we obtain again a different order of features' importance for the both models. In nIVR, the order was dew point rain rate and humidity; instead in IVR only humidity and rain rate were important. Figure 16 shows this graphically.

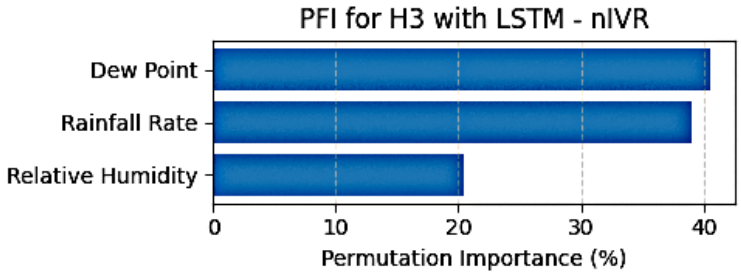


(**a**) PFI for *H*2 applied to the ANN model - nIVR case

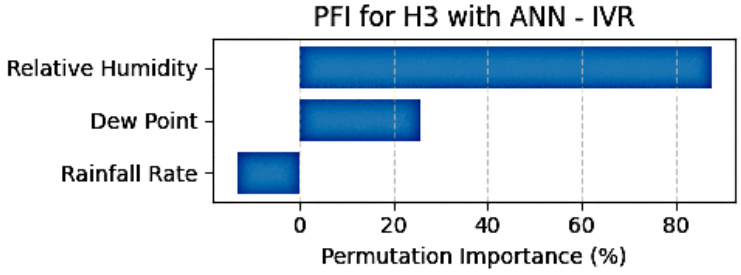


(**b**) PFI for *H*2 applied to the ANN model - IVR case

Figure 15. Comparison of the result of permutation feature importance for *H*2 applied to predicted values.



(**a**) PFI for *H*3 applied to LSTM model - nIVR case



(**b**) PFI for *H*3 applied to the ANN model - IVR case

Figure 16. Comparison of the result of permutation feature importance for *H*3 applied to predicted values.

6. Discussion

As presented in our previous work [30], in this study, we analyzed the results from two different points of view. On the one hand, we compared our application cases to related works in the literature. In particular, we focused on a very recent literature review presented in [27], where the authors analyzed 187 primary studies (from 2002 to 2023) that applied neural network techniques for forecasting changes in groundwater levels. From this article, we can see that LSTM is a technique widely used for this type of study, yielding promising and novel results. Additionally, the most common influencing factors used in model construction involved weather variables such as humidity, precipitation, temperature, etc., water table levels (from piezometers) and aquifer recharge. In our case, the weather variables were selected from previous studies showing their specific influence on the water table levels. Furthermore, the recharge was obtained from the river discharge. Finally, the article applied MSE, RMSE, and R^2 as the most commonly used performance metrics.

On the other hand, as with any other methodology that proposes software reuse, it was necessary to evaluate the development of reusable components. This means we need to analyze the time required to develop the services of the reusable components (reusable components are pre-built pieces of software that can be used in multiple applications). In this case, we applied the same methodology used in [30], in which we analyzed the services needed to implement each activity and measured the time it takes for each of them to be resolved (in number of days). A service is a task within each big data process activity defined for each application case. In Figure 17, we show the time required for the two application cases defined as reusable: RC prediction non-irrigation VR and RC prediction irrigation VR. As we can see, the services are numerated and represent specific activities. For instance, *S*0 is the service that determines the domain-case problem, for which more than 30 days were needed in the case RC prediction non-irrigation VR. This is because we held meetings with experts, determined new needs, reviewed stored data, drew conclusions, and formulated hypotheses. For the second application case, this same service took only 2 days to define the scope with experts and analyze the work to be done.

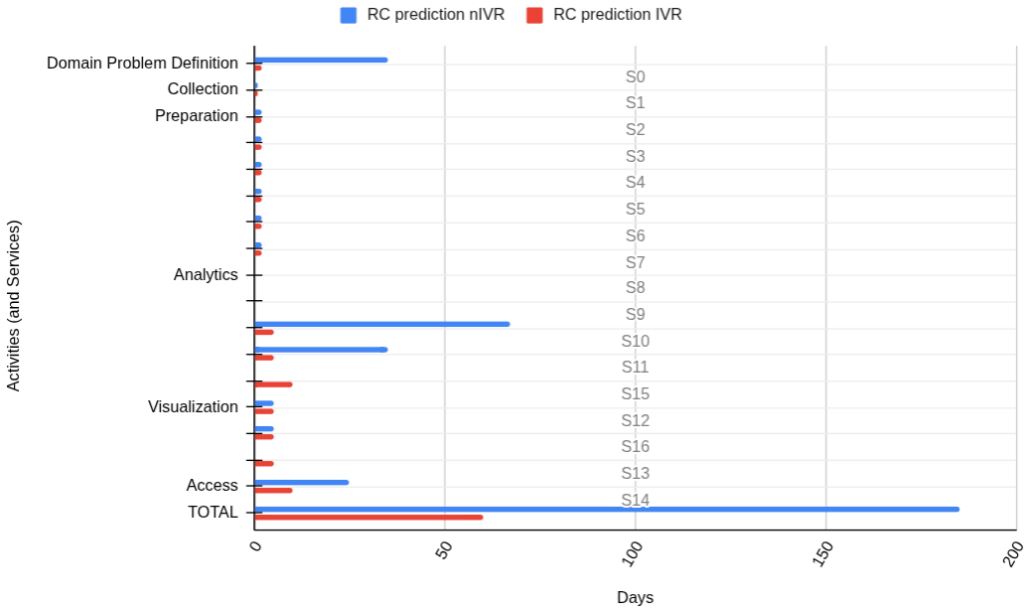


Figure 17. Time required (in days) for the development of activities for the two application cases.

Another example is shown with the services of collection and preparation activities. We can see that these services, which include *S*3 (Imputation of null values), *S*4 (Filtering

piezometers), *S*7 (Creating unified datasets), etc., took very little time because the source and prepared datasets were retrieved from CoVaMAT, and only analyzed for these two application cases. In contrast, services *S*10 (Computing hyperparameter and predicting river-water table) and *S*11 (Computing hyperparameter and predicting weather-water table) took more than two months to complete, as they were developed entirely from scratch. For the second application case (RC prediction irrigation VR), the same two services were fully developed in just 10 days because only minor adjustments and configuration changes were needed.

An additional aspect can be observed in services *S*7 (Computing a correlation analysis of river discharges—water table levels) and *S*8 (Computing a correlation analysis weather variables—water table levels). These services have no time assigned because they were performed in the cases already stored: DC influences nIVR and DC influences nIVR.

In total, we can see that the second application case, which reused previous stored cases, including RC prediction non-irrigation VR, achieved a percentage gain of approximately 208.33% (from 60 to 185).

6.1. Threats to Validity

6.1.1. From the Analytics

Firstly, LSTM is a specialized type of RNN designed to address the limitations of traditional RNNs, particularly their inability to capture long-term dependencies. However, despite its wide-spread use, some challenges related to LSTM initialization and optimization persist, as reported in [31]. Some common pitfalls include overfitting, which occurs when a model learns the training data too well, including noise and outliers, and fails to generalize to unseen data. We actually tested several configurations by changing optimization functions such as tanh, dropout rate, etc. By doing this, the three configurations shown in the first case; for example, allowed us to generalize our tests and provide the combinations that gave us the best results. With other combinations, we observed (1) significant dispersion in performance metrics when we trained the model multiple times, and/or (2) results with poor model performance. At the same time, we also tested various configurations of other predictive algorithms such as ANN and SVM, considered as the most widely used techniques in groundwater level predictions [27,32]. However, not all possible combinations were tested, and similarly, other techniques or temporal architectures were not tested either.

Another pitfall to consider is the appropriate handling of preprocessing and feature analysis. In our approach, feature analysis is considered an intrinsic part of the method during top-down variety identification. Therefore, we mitigate this threat in two ways: by incorporating domain analysis into the method and by reusing domain assets that have proven useful in analyzing similar situations, albeit in different contexts. Furthermore, the preprocessing variety is also stored in CoVaMaT, allowing the reuse of databases that have already been prepared and transformed for the problem domain.

When working with ML models, outliers and null values can significantly degrade performance, leading to various proposals for extending and/or improving the models. In our work, we reused preprocessed datasets where these values had been imputed, and the focus of the case studies was on obtaining the best performance from the available datasets. By generalizing this approach, any variation in the datasets used may lead to a review of the techniques employed, both for handling outliers and null values, as well as for performing the prediction itself.

6.1.2. From the Process to the Problem

The case studies were developed by members of the research team, which poses a risk to their application by third parties. In this case, the learning curve for transferring this approach to industry could represent an investment that any company should carefully consider before proceeding. To mitigate this risk, our process for identifying and building reusable big data systems relies on easily identifiable steps; however, this does not eliminate the need to develop a procedural guide for their transfer.

Bridging the gap between industry needs and new methodological approaches is essential for the successful application of our approach, which is important for any new proposal focused on software development [33]. In our case, the risk was mitigated by developing the conceptual part of the approach in conjunction with its application in an industrial setting, specifically to address precision agriculture problems. However, generalizing and scaling the application to other domains and a wider range of problems remain to be done; this again highlights the need to delve deeper into the process model and the measurement of reuse as key adoption factors.

The results obtained are promising; however, the analysis is based on measurements from 15 piezometers within a 2 km radius of the river, while relevant for domain experts, the results could vary if a larger number of measurements and other variables that could affect groundwater recharge were considered. The study is preliminary in this respect but useful for demonstrating the contextual influence and the feasibility of building reusable big data systems.

Based on the conclusions drawn from CoVaMaT, it was observed that, during the irrigation period, river flow and weather variables shown low correlation with groundwater levels. However, ML models showed that the inclusion of these variables was useful for prediction, demonstrating very good model performance. Recent studies, such as [34,35], have also shown that, variables with low linear correlation contributed significantly to model accuracy in LSTM models; that is, groundwater forecasting also benefited from including variables that shown low instantaneous correlation with groundwater levels. This is consistent with the ability of LSTM networks to capture nonlinear and temporal dependencies.

Analyzing groundwater level variations is complex, and limiting it to the relationship with river flow is a first approach to identifying variations while keeping other factors constant. For example, variations in terrain elevation and potential runoff diversions near or far from a piezometer can influence level variations, as can soil type and land use (cropland, grazing, urban, etc.). This means that the scope of the contextual analysis must be controlled, managing complexity while including the most relevant factors in the context under study. In our case, for instance, soil type was excluded from the analysis because the area where the piezometers are located is practically homogeneous. According to official soil classifications, the study area corresponds to a recent floodplain zone, which is characterized by young and moderately deep soils that exhibit rapid permeability. Topography was also not included as a predictive variable, as the study area is predominantly flat, with minimal elevation differences among piezometer locations, and therefore does not introduce significant spatial variability affecting water table levels.

Finally, our experience shows that even when data analysis is performed multiple times to reduce errors and improve performance, the best result in one context may not be the best in another. This opens up two perspectives in analytics processing: first, every result is relative to the context in which the analysis was applied as we could see from the different best model for each case and hypothesis (Section 5); and second, the results and the process itself can be reused in similar contexts, as we could see from our proposal's instantiation (Section 4). Therefore, the best analyzes will not only come from

the most comprehensive measurements of the results but from identifying similarities in the addressed problems. Identifying reuse possibilities can lead to a better understanding, and more accurate and more efficient solutions in terms of effort, as shown Figure 17.

The different types of varieties in big data systems are interrelated. Although many data analytics techniques (processing variety) are generally applicable, deeper processing depends on the type of problem (context variety); for example, the unique characteristics of wildfire prediction indicate that addressing spatial heterogeneity is essential to avoid biased predictions [36]. As the authors suggest, this area should explore the use of graph-based models and domain adaptation, clearly linking processing and problem knowledge.

7. Conclusions

Analyzing the factors that influence groundwater recharge and predicting its fluctuations is a complex process that requires in-depth knowledge of both data science and the problem domain, as we have highlighted in this article. Improvement in the results in hydrological systems analysis can stem from a combination of systematically linked algorithms and knowledge. Undoubtedly, understanding the hypotheses of data analysis is essential for making predictions, correlations, and so on, leading to accurate results. However, performing this analysis and architecting software, while making the procedure reproducible across diverse contexts, requires a comprehensive approach that involves both engineering and data science. This paper shows a development approach in which both aspects are combined into a procedure that considers how the variety of the context can influence the analytical results.

Despite obtaining promising results, we have listed some threats and considerations that would allow for a better understanding of the problem and a more in-depth analysis in future work. For instance, future extensions of our approach will consider the inclusion of standards for documenting contextual variety, such as WaterML (https://www.ogc.org/standards/waterml/ accessed on 1 February 2026) or INSPIRE (Infrastructure for Spatial Information in Europe (https://inspire-mif.github.io/technical-guidelines/data/hy/dataspecification_hy.html accessed on 1 February 2026)) for documenting contextual variety.

In this regard, in addition to extensions that allow for the standardization of domain vocabulary, the use of hybrid models combining elements or techniques from two or more distinct models or methodologies is also an aspect to explore in the search for better performance of the results.

Similarly, and considering the analytical-context combination, expanding the scope of the problem to include; for example, aspects of flooding in the terrain could broaden the analysis of influences, since these aspects can be seen as a recharge mechanism. In this regard, one of the current challenges is implementing a mechanism to measure the water level in the flooded area near each piezometer. Additionally, the irrigation canal network is another factor that can be considered in the influence of groundwater level variations in the study area. These influences could stem not only from irrigation itself, but also from how water flow is used to fill the primary and secondary canals.

Our future work is therefore focused both on expanding the scope of the study to include new potential effects on groundwater aquifer recharge, and on improving the context-based big data system development method and its supporting tool, CoVaMaT.

Author Contributions: Conceptualization, A.B. and A.C.; investigation, A.B. and A.C.; methodology, A.B. and A.C.; software, W.G.; writing—original draft preparation, A.B.; writing—review and editing, A.C.; supervision, A.B.; data providing and resources, A.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the articlel. Further inquiries can be directed to the corresponding author.

Acknowledgments: This work is partially supported by the UNComa project 04/F019 "Variety modeling in Big Data Systems" 2022–2026.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BDS	Big Data Systems
SPL	Software Product Line
OVM	Orthogonal Variability Model
CoVaMaT	Context-based Variety Management Tool
ML	Machine Learning
DC	Domian Case
RC	Reusable Case
VR	Villa Regina
IVR	Irrigation Villa Regina
nIVR	Non-Irrigation Villa Regina
LSTM	Long Short-Term Memory
ANN	Artificial Neural Network
SVM	Support Vector Machine
R^2	Coefficient of determination
RMSE	Root Mean Square Error
MSE	Mean Square Error
MAE	Mean Absolute Error

References

1. Bahga, A.; Madisetti, V. *Big Data Science & Analytics: A Hands-On Approach*, 1st ed.; VPT: Atlanta, GA, USA, 2016.
2. Erl, T.; Khattak, W.; Buhler, P. *Big Data Fundamentals: Concepts, Drivers & Techniques*, 1st ed.; Prentice Hall Press: Upper Saddle River, NJ, USA , 2016.
3. Klein, J.; Buglak, R.; Blockow, D.; Wuttke, T.; Cooper, B. A reference architecture for big data systems in the national security domain. In *Proceedings of the 2016 IEEE/ACM 2nd International Workshop on Big Data Software Engineering (BIGDSE), Austin, TX, USA, 16 May 2016*; IEEE: Piscataway, NJ, USA, 2016; pp. 51–57.
4. Davoudian, A.; Liu, M. Big Data Systems: A Software Engineering Perspective. *ACM Comput. Surv.* **2020**, *53*, 1–39. [CrossRef]
5. Borrison, R.; Klöpper, B.; Chioua, M.; Dix, M.; Sprick, B. Reusable Big Data System for Industrial Data Mining—A Case Study on Anomaly Detection in Chemical Plants. In *Proceedings of the Intelligent Data Engineering and Automated Learning—IDEAL 2018, Madrid, Spain, 21–23 November 2018*; Yin, H., Camacho, D., Novais, P., Tallón-Ballesteros, A.J., Eds.; Springer: Cham, Switzerland, 2018; pp. 611–622.
6. Epperson, W.; Wang, A.Y.; DeLine, R.; Drucker, S.M. Strategies for Reuse and Sharing among Data Scientists in Software Teams. In *Proceedings of the ICSE-SEIP '22, Pittsburgh, PA, USA, 22–24 May 2022*; Association for Computing Machinery: New York, NY, USA, 2022. [CrossRef]
7. Garrido, W.; Buccella, A.; Cechich, A.; Montenegro, A. Análisis de influencias en la recarga de las napas freáticas: Un caso de estudio en reusabilidad contextual de sistemas big data. *JAIIO Jorn. Argent. Inform.* **2025**, *11*, 71–84.
8. Muhuri, P.S.; Chatterjee, P.; Yuan, X.; Roy, K.; Esterline, A. Using a Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) to Classify Network Attacks. *Information* **2020**, *11*, 243. [CrossRef]
9. Pasquetto, I.; Randles, B.; Borgman, C. On the Reuse of Scientific Data. *Data Sci. J.* **2017**, *16* , 1–9. [CrossRef]
10. Custers, B.; Uršič, H. Big data and data reuse: A taxonomy of data reuse for balancing big data benefits and personal data protection. *Int. Data Priv. Law* **2016**, *6*, 4–15. [CrossRef]

11. Xie, Z.; Chen, Y.; Speer, J.; Walters, T.; Tarazaga, P.A.; Kasarda, M. Towards Use And Reuse Driven Big Data Management. In *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries, New York, NY, USA, 7–16 June 2009*; Association for Computing Machinery: New York, NY, USA, 2015; pp. 65–74.
12. Klein, J. Reference Architectures for Big Data Systems, Carnegie Mellon University's Software Engineering Institute Blog. 2017. Available online: http://insights.sei.cmu.edu/blog/reference-architectures-for-big-data-systems/ (accessed on 9 June 2021).
13. Nadal, S.; Herrero, V.; Romero, O.; Abelló, A.; Franch, X.; Vansummeren, S.; Valerio, D. A software reference architecture for semantic-aware Big Data systems. *Inf. Softw. Technol.* **2017**, *90*, 75–92. [CrossRef]
14. Cuesta, C.E.; Martínez-Prieto, M.A.; Fernández, J.D. Towards an Architecture for Managing Big Semantic Data in Real-Time. In *Proceedings of the Software Architecture*; Drira, K., Ed.; Springer: Berlin/Heidelberg, Germany, 2013; pp. 45–53.
15. Duggan, J.; Elmore, A.J.; Stonebraker, M.; Balazinska, M.; Howe, B.; Kepner, J.; Madden, S.; Maier, D.; Mattson, T.; Zdonik, S. The BigDAWG polystore system. *ACM SIGMOD Rec.* **2015**, *44*, 11–16. [CrossRef]
16. Pohl, K.; Böckle, G.; Linden, F. *Software Product Line Engineering: Foundations, Principles and Techniques*; Springer: Berlin/Heidelberg, Germany, 2005.
17. Osycka, L.; Cechich, A.; Buccella, A.; Montenegro, A.; Muñoz, A. CoVaMaT: Functionality for Variety Reuse through a Supporting Tool. In *Proceedings of the Cloud Computing, Big Data & Emerging Topics*; Springer Nature: Cham, Switzerland, 2023; pp. 57–74.
18. Blanken, P. *Essentials of Water: Water in the Earth's Physical and Biological Environments*; Cambridge University Press: Cambridge, UK, 2024.
19. Gebreslassie, H.; Berhane, G.; Gebreyohannes, T.; Hagos, M.; Hussien, A.; Walraevens, K. Water Harvesting and Groundwater Recharge: A Comprehensive Review and Synthesis of Current Practices. *Water* **2025**, *17*, 976. [CrossRef]
20. Lallahem, S.; Mania, J.; Hani, A.; Najjar, Y. On the use of neural networks to evaluate groundwater levels in fractured media. *J. Hydrol.* **2005**, *307*, 92–111. [CrossRef]
21. Djurovic, N.; Domazet, M.; Stricevic, R.; Pocuca, V.; Spalevic, V.; Pivic, R.; Gregoric, E.; Domazet, U. Comparison of Groundwater Level Models Based on Artificial Neural Networks and ANFIS. *Sci. World J.* **2015**, *2015*, 13. [CrossRef] [PubMed]
22. Shamsuddin, M.K.N.; Mohd Kusin, F.; Sulaiman, W.; Ramli, M.; Tajul Baharuddin, M.F.; Adnan, M.S. Forecasting of Groundwater Level using Artificial Neural Network by incorporating river recharge and river bank infiltration. *MATEC Web Conf.* **2017**, *103*, 04007. [CrossRef]
23. Liu, Q.; Gui, D.; Zhang, L.; Niu, J.; Dai, H.; Wei, G.; Hu, B.X. Simulation of regional groundwater levels in arid regions using interpretable machine learning models. *Sci. Total Environ.* **2022**, *831*, 154902. [CrossRef] [PubMed]
24. Eftekhari, M.; Khashei-Siuki, A. Evaluating machine learning methods for predicting groundwater fluctuations using GRACE satellite in arid and semi-arid regions. *J. Groundw. Sci. Eng.* **2025**, *13*, 5–21. [CrossRef]
25. Yang, Y.; Zhao, J. Forecasting the Spatio-Temporal Evolution of Groundwater Vulnerability: A Coupled Time-Series and Hydrogeological Modeling Approach. *Water* **2025**, *17*, 3033. [CrossRef]
26. Baki, A.M.; Ghavami, S.M. A modified DRASTIC model for groundwater vulnerability assessment using connecting path and analytic hierarchy process methods. *Environ. Sci. Pollut. Res.* **2023**, *30*, 111270–111283. [CrossRef] [PubMed]
27. Afful, S.K.; Boateng, C.D.; Ahene, E.; Aryee, J.N.A.; Wemegah, D.D.; Gidigasu, S.S.R.; Britwum, A.; Osei, M.A.; Gilbert, J.; Touré, H.; et al. A systematic review of neural network applications for groundwater level prediction. *Discov. Appl. Sci.* **2025**, *7*, 942. [CrossRef]
28. Jesse, G.; Boateng, C.D.; Aryee, J.N.; Osei, M.A.; Wemegah, D.D.; Gidigasu, S.S.; Britwum, A.; Afful, S.K.; Touré, H.; Mensah, V.; et al. A systematic review of machine learning models for groundwater level prediction. *Appl. Comput. Geosci.* **2025**, *28*, 100303. [CrossRef]
29. Mandler, H.; Weigand, B. A review and benchmark of feature importance methods for neural networks. *ACM Comput. Surv.* **2024**, *56*, 1–30. [CrossRef]
30. Buccella, A.; Cechich, A.; Saurin, F.; Montenegro, A.; Rodríguez, A.; Muñoz, A. A Context-Based Perspective on Frost Analysis in Reuse-Oriented Big Data-System Developments. *Information* **2024**, *15*, 661. [CrossRef]
31. Al-Selwi, S.M.; Hassan, M.F.; Abdulkadir, S.J.; Muneer, A.; Sumiea, E.H.; Alqushaibi, A.; Ragab, M.G. RNN-LSTM: From applications to modeling techniques and beyond—Systematic review. *J. King Saud Univ.–Comput. Inf. Sci.* **2024**, *36*, 102068. [CrossRef]
32. Sakovich, N.; Aksenov, D.; Pleshakova, E.; Gataullin, S. A neural operator using dynamic mode decomposition analysis to approximate partial differential equations. *AIMS Math.* **2025**, *10*, 22432–22444. [CrossRef]
33. Daun, M.; Brings, J.; Aluko Obe, P.; Tenbergen, B. An industry survey on approaches, success factors, and barriers for technology transfer in software engineering. *Softw. Pract. Exp.* **2023**, *53*, 1496–1524. [CrossRef]
34. Mauladdawilah, H.; Balfaqih, M.; Balfagih, Z.; Pegalajar, M.d.C.; Gago, E.J. Deep Feature Selection of Meteorological Variables for LSTM-Based PV Power Forecasting in High-Dimensional Time-Series Data. *Algorithms* **2025**, *18*, 496. [CrossRef]

35. Epting, J.; Huggenberger, P.; Radny, D.; Hammes, F.; Hollender, J.; Page, R.M.; Weber, S.; Bänninger, D.; Auckenthaler, A. Spatiotemporal scales of river-groundwater interaction—The role of local interaction processes and regional groundwater regimes. *Sci. Total Environ.* **2018**, *618*, 1224–1243. [CrossRef] [PubMed]
36. Ejaz, N.; Choudhury, S. A comprehensive survey of the machine learning pipeline for wildfire risk prediction and assessment. *Ecol. Inform.* **2025**, *90*, 103325. [CrossRef]