Ref. Ares(2021)4901648 - 01/08/2021

European Commission

HORIZON 2020

PROJECT DELIVERABLE REPORT

NAIADES

Greening the economy in line with the sustainable development goals

D5.9 Predictive AI analytics for the quality of the water – Mid-term

A holistic water ecosystem for digitisation of urban water sector

SC5-11-2018

Digital solutions for water: linking the physical and digital world for water solutions

"This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 820985"

Document Information

Grant Agreement Number	820985	Acronym	NAIADES		
Full Title	A holistic water ecosystem for digitization of urban water sector				
Topic	SC5-11-2018: Digital solutions for water: linking the physical and digital world for water solutions				
Funding scheme	IA - Innovation Action				
Start Date	1stJUNE 2019		**Duration**	36 months	
Project URL	www.NAIADES-project.eu				
EU Project Officer	Alexandre VACHER				
Project Coordinator	CENTER FOR RESEARCH AND TECHNOLOGY HELLAS - CERTH				
Deliverable	D5.9 Predictive AI analytics for the quality of the water – Mid-term				
Work Package	WP5 – Smart Framework: AI analytics and predictive services				
Date of Delivery	**Contractual**	M18	**Actual**	M17	
Nature	R - Report		**Dissemination Level**	PU-PUBLIC	
Lead Beneficiary	AIMEN				
Responsible Author	Dr. Juan Manuel Fernández Montenegro		Email	Juan.fernandez@aimen.es	
			Phone	+34697 99 14 97	
Reviewer(s):	Guardtime, DISY				
Keywords	AI, water quality, forecasting, short-term				

Revision History

Version	Date	Responsible	Description/Remarks/Reason for changes
0.1	09/10/2020	AIMEN	Table of Contents
0.2		AIMEN	Report write-up
	16/11/2020	DISY	Internal Review
0.3	18/11/2020	AIMEN	Inclusion DISY's Internal Review
	18/11/2020	GT	Internal Review
0.4	18/11/2020	GT, AIMEN	Inclusion of GT's Internal Review
1.0	30/11/2020		Review and Release
1.1	11/05/2021	AIMEN	PO review applied
	11/05/2021	GT	Internal Review

2.0	21/05/21	AIMEN	Review and Release

Contents

Abbreviations

AI	Artificial Intelligence
AWCS	Awareness and Behavioural Change Support
DCAs	Data Collector/Aggregators
DM	Data Manager
DSS	Decision Support System
HMI	Human Machine Interface
LSTM	Long Short-Term Memory
RNN	Recurrent Neural Network
SVR	Support Vector Regression
WTP	Water Treatment Plant

List of Tables

List of Figures

1 Summary

This report focuses on the description of the procedures and first prototype being developed on Task 5.5 – Water Quality Forecasting Tool. This mid-term deliverable contains information that describes the purpose of the tool and the elements that will compose it, this is, its design, its functioning and all the extra features that are required to be part of NAIADES platform. It also describes all the requirements necessary for the tool to perform positively, being the availability of quality real data the principal one.

2 Introduction

Water sources generally require treatment prior to consumption/usage to ensure that they do not present a health risk to the user. Health risks from poor quality water will be often due to microbiological or chemical contamination.

Water quality in artificial and natural courses, such as pipes, canals, streams and rivers, depends on a wide broad of aspects and it needs to be characterized periodically for water treatment: pathogens, pH, conductivity, suspended solids, organic matter, etc. Most of these parameters are usually determined by a laboratory. As these take time, they are sometimes obtained too late to be useful to make a decision.

Prediction with water quality models could allow a quick response to extreme events that change quality characteristics in a short period of time, such a high organic load (i.e. algae blooms) or suspended solids (i.e. storms). They can provide real time water characterization thus removing the necessity of slow laboratory measurements. Water quality modelling involves the prediction of water pollution using mathematical simulation techniques. A typical water quality model consists of a collection of formulations representing physical mechanisms that determine position and momentum of pollutants in a water body. As there are many competing numerical models, it becomes difficult to evaluate all these models and this can be automated by using methods from the field of machine learning.

The goal of the NAIADES AI analytics for water prediction service is the provision of future water quality parameters so the system is able to take an action regarding events related with the quality of water in a dynamical and active data driven approach and to increase the performance and safety of the current system.

The next subsections present the aims and objectives of the water prediction solution for NAIADES platform and for the specific necessities of NAIADES end users (use cases). The rest of the document is distributed as follows. Section 3 locates and explains the tool position within NAIADES platform architecture. Section 4 describes the internal functioning of the tool and the requirements for its integrations within the NAIADES cloud platform. Section 5 summarises the following steps after the first prototype of the tool is ready and integrated and Section 6 shows the conclusions.

2.1 Purpose inside NAIADES

NAIADES water quality forecast aims to provide short-term future predictions of water quality related parameters such as pH, chlorine and turbidity at a given location. The predictions will take into account different sets of variables, starting for the parameters' evolution (historical data), the parameters combined with weather information and the parameters combined with other parameters measured at the same location. The expected outcomes will depend on the necessities of the end users, aiming to provide up to seven days' worth of hourly forecasts. The water quality service could be applied at any location of the water network. It has been required to cover two necessities from the different end users.

2.2 Use Cases

NAIADES end users presented different problems/processes that they would like to overcome/improve using NAIADES platform. Water quality forecast service was chosen as (partial) solution for two problems related to water quality in public fountains used for bathing during summer and water treatments optimization on drinking Water Treatment Plants (WTP).

Figure 1. Images depicting NAIADES use cases that require water quality forecast. The image on the left represents drinking water for the Treatments suggestions use case. The image on the right shows Carouge Fontaine des Tours.

2.2.1 UCC2 Fountains

The use case proposed at Carouge pilot aims to ameliorate the management of the city fountains for three main reasons.

1. Reduce the amount of time the city staff have to spend for water quality monitoring labour.
2. Improve maintenance activities planning so they do not disrupt the fountain service.
3. Guarantee water quality.

Water Quality forecast provides a predictive maintenance service; it helps to solve all the three points by providing advanced predictions so the staff can organize their work beforehand; by giving hints at least two days in advance about a possible deterioration of water quality that could end in fountain maintenances. According to those hints, the staff can act to plan the maintenance in advance so it does not disturb the fountain service; and all of this allows a better planning of the treatments to keep always the optimal water quality.

2.2.2 UClsWTP

This use case's main objective is to ensure the distribution network will always have the best possible water quality coming from the drinking WTP. It was initially demanded by Braila pilot but since they cannot provide the required data for modelling due to confidentiality issues, a laboratory pilot was designed to showcase the proposed solution. NAIADES proposes a service that optimizes the WTP treatments, ensuring real-time water quality monitoring and treatments that adapt to any change on the quality of the inlet water.

Water Quality Forecast service comes as an extension of the treatments service by providing future inlet water quality forecasts. These predictions are to be used by the treatments service to generate suggestions about the optimal treatments for the predicted water quality. As a result, WTP staff can take into account the future suggestions for planning of treatments and required stock.

3 Water Quality Forecast Tool inside NAIADES architecture

NAIADES architecture described in NAIADES project's deliverable 2.9 presents Water Quality Forecast Tool as one of their AI services on the Application layer (see Figure 2).

The AI services are part of the Cloud Platform and all their functionality is hold within the Cloud Platform. Their source of data is the Data Manager (DM), which is responsible of distributing all the information required by NAIADES tools. NAIADES AI services will not connect to external data sources, all their data depends on NAIADES platform. This data will come from the Data Collector/Aggregators (DCAs), the decision Support System (DSS), the Human Machine Interface (HMI), the Awareness and Behavioural Change Support (AWCS) or other AI services. The data required by the Water Quality Forecast Tool will arrive to the DM from the DCAs, which will collect the data from quality sensors and local weather data stations, and from the Weather Prediction service. NAIADES DM has two submodules, one manages real-time data (freshly uploaded to the platform) and the other for historical data (time-series data). The prediction tool will access the historical submodule to get the data to create the prediction model and the real-time one to provide the forecasts (there is the possibility that the historical one is required depending on the machine learning algorithm finally used – see section 6).

The outcomes from the AI service will go to the DM, although, they have to go through the Data Validation Module first. The outcomes need to be formatted, following the NAIADES data models, and signed before being sent to the Data Model Validation Module. This module will check if the format fits NAIADES specifications, forwarding it to the DM if everything is correct or returning an error if not. The NAIADES data models for Water Quality and other related information regarding the integration of the Water Quality Forecast Tool data is shown in section 4.3.

The information provided by the Water Quality Forecast Tool can be used by any other NAIADES service (AI, DSS, AWCS) or displayed on the HMI or other dedicated App. Furthermore, NAIADES marketplace (more info in NAIADES deliverable D7.4) will present the tool description and its functionalities to any potential user.

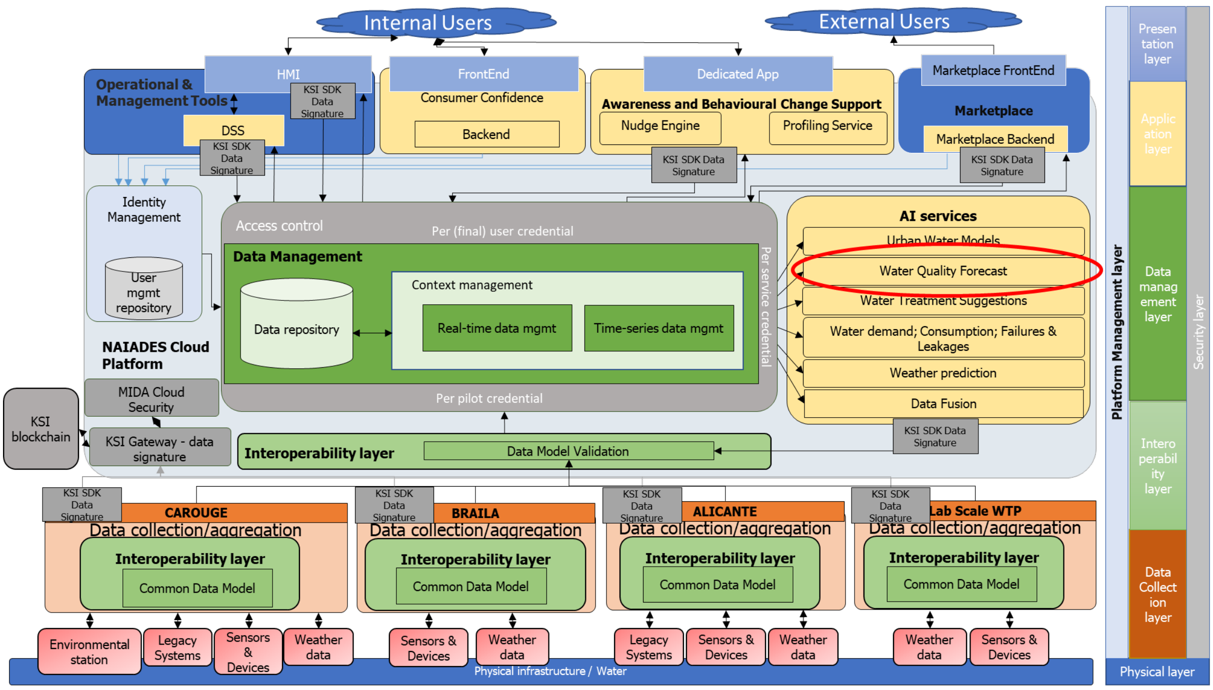

Figure 2. Water Quality Forecast Tool in NAIADES Platform

4 Water Quality Forecast Tool insights

4.1 AI modelling

The Water Quality Forecast Tool is mainly a machine learning algorithm that analyses thousands of historical values of water quality and weather parameters to understand their behaviour and provide estimations of future values[1]. The machine learning approach allows the modelling of complex water processes without the use numerical models describing each subprocess, thus simplifying the modelling task. Machine learning algorithms usually requires three stages to be successfully deployed: training, validation and testing. The next subsections will describe which data is going to be used, the training process, and the validation and testing stages.

4.1.1 Data acquisition

The main sources of data of this tool are the water quality sensors information and the weather data. The weather data will be always required but the quality parameters to be measured will depend on the data provided by each use case (see section 2.2). The quality parameters are shown in Table 1 presented by use case. These parameters will be collected by quality sensors, installed at the pilots, in an hourly basis or at least twice a day, depending on the predictions' frequency required by the use cases. The predictions frequency cannot be higher than the acquisition one.

Water Quality Parameters	UCC2 Fountains	UClsWTP	Source
Free Chlorine	✔	✔	Sensor
Total Chlorine	✔	✔	Sensor
pH	✔	✗	Sensor
Redox	✔	✗	Sensor
Turbidity	✔	✔	Sensor
Temperature	✗	✔	Sensor
Conductivity	✗	✔	Sensor

Table 1. Water Quality Parameters

The weather parameters are presented in Table 2. The weather data will be collected from two sources. The hourly monitored data will be uploaded to the platform at least once a day from the local weather stations and they will be used for immediate predictions (next hour forecasts). The NAIADES weather predictions service will provide four forecasts per day at least of the following two days. These forecasts will be used to provide future water quality forecasts.

[1] Hafeez, S., Wong, M. S., Ho, H. C., Nazeer, M., Nichol, J., Abbas, S., ... & Pun, L. (2019). Comparison of machine learning algorithms for retrieval of water quality indicators in case-II waters: a case study of Hong Kong. *Remote sensing*, *11*(6), 617.

Weather Data	**Source**
Precipitation	Local Weather Station
Temperature	Local Weather Station
Precipitation Forecast	Weather Prediction Tool
Temperature Forecast	Weather Prediction Tool

Table 2. Input Weather Parameters

Machine learning modelling usually requires a huge set of data, being required thousands of samples for methods such as deep learning algorithms, furthermore, for processes that changes seasonally, it is preferably to have years of data to be able to generate robust models, otherwise, it will be required to re-learn data in regular intervals or dividing the data is sets where dynamic changes were detected. Since time series data are required for training, the DM historical submodule will be used to obtain as much data as possible. The dataset obtained is then divided in two sets for training and validation, in an 80-20 proportion.

4.1.2 Training

The training phase of the modelling process requires the largest amount of data. It will take at least eighty percent of all the historical data as input of the machine learning algorithms; being required afterwards an exhaustive testing using new data to check and avoid overfitting problems.

Three different algorithms will be used for water quality modelling: Support Vector Regression (SVR) and two deep learning approaches. SVR is the first algorithm to be used since it does not require huge amount of data to work reliably. This algorithm receives as input a time series of the parameters of interest (water quality parameters excepting the one to predict) and their input date; and it also receives separately the time series of the parameter to be predicted. The algorithm calculates the best function that defines the relationship between both time series; this will be the model. This approach is the simplest and it is not expected to provide the best results.

The deep learning approaches will be implemented once the amount of data is considerable. The first deep learning solution will be similar to the SVR solution. The same input time series will be used as input to a simple neural network so the network parameters are adjusted to the training data. The second deep learning solution will take into account the last variations of the data in the time series. Recurrent Neural Networks (RNN) with Long Short-Term Memory (LSTM) blocks takes into account previous values of each variable when it tries to find a relation with the parameters to be predicted[2]. Therefore, the RNN will take as input the previous values or a time series of 'times series'; and these previous values are linked to the next value of the variable to be predicted (see Table 3). The RNN using time series of 'time series' will imply that the tool will only require to use the historical data base.

[2] Qi, C., Huang, S., & Wang, X. (2020). Monitoring Water Quality Parameters of Taihu Lake Based on Remote Sensing Images and LSTM-RNN. *IEEE Access*, *8*, 188068-188081.

SVR and Deep Learning Solution 1				Deep Learning Solution 2 (RNN)			
Measured Variables			Measured Variable to be predicted	Measured Variables			Measured Variable to be predicted
pH t-5	Turbidity t-5	Time t-5	Cl t-5	pH t-11..t-6	Turbidity t-11..t-6	Cl t-11..t-6	Cl t-5
pH t-4	Turbidity t-4	Time t-4	Cl t-4	pH t-10..t-5	Turbidity t-10..t-5	Cl t-10..t-5	Cl t-4
pH t-3	Turbidity t-3	Time t-3	Cl t-3	pH t-9..t-4	Turbidity t-9..t-4	Cl t-9..t-4	Cl t-3
pH t-2	Turbidity t-2	Time t-2	Cl t-2	pH t-8..t-3	Turbidity t-8..t-3	Cl t-8..t-3	Cl t-2
pH t-1	Turbidity t-1	Time t-1	Cl t-1	pH t-7..t-2	Turbidity t-7..t-2	Cl t-7..t-2	Cl t-1
pH t	Turbidity t	Time t	Cl t	pH t-6..t-1	Turbidity t-6..t-1	Cl t-6..t-1	Cl t

Table 3. Example of Input data for SVR and Deep Learning Solution 1 vs RNN. The RNN solution can also be tested with the previous value instead of using the 'time series'

Only the water quality parameters are being used for the first trials. Currently, there is not data available from the use cases to train this tool, thus, data freely available on the Internet is being used; and it is difficult to find water quality databases together with the weather information. The weather data will be added when the use cases are able to provide all the required information.

4.1.3 Validation and testing

The training stage returns a model that is fitted to show the best performance for the training data. The validation and testing stages function is to improve the performance with new data and to probe the model works properly with any incoming data[3].

The validation stage takes part together with the training. Twenty percent of the historical data is usually used for validation purposes. The data is used to evaluate the training model changing the parameters of the model, so it is not so biased by the training data.

Finally, the test stage is done once a final version of the model is created. The test stage uses completely new data to evaluate the model. The accuracy obtained with the test data will be the most representative of the model performance. The final model will be the one that probes the greater performance with the test data and it will be stored in a file to be used by the inference tool.

[3] Mitchell, T. (1997). Introduction to machine learning. *Machine Learning*, *7*, 2-5.

4.2 Model inference

Machine learning inference refers to the use of the final model with real data. The next subsections describe how the data must be acquired and how it must be used with the model.

4.2.1 Data acquisition

It has already been presented in section 4.1.2 that the data should be collected from the historical database (DM time series data submodule) for training purposes. Nevertheless, once the model is ready, there are two possibilities depending on the final machine learning algorithm used.

- The DM real time data submodule will be used if the training algorithm is SVR, the first deep learning solution or the RNN using only the instant values per variable.
- The DM time series data submodule will be used if the RNN that uses time series of 'time series' is the prediction algorithm.

The final decision will be made once all the algorithms are evaluated in terms of accuracy and computation requirements.

4.2.2 Inference

Once the data is retrieved from the DM it should be redistributed so to fit the required inputs of the model (as shown in 4.1.1). These inputs will feed the inference tool (Water Quality Forecast Tool). The inference tool is in charge of using the prediction algorithm in combination with the final model; this is it runs the algorithm with the parameters stored in the model to predict the desired values. During the training phase, a model will be created per water quality parameter. Each model will be used by the tool for predicting the respective parameters to be forecasted. For example, it will provide 4 prediction per day of pH and chlorine predictions of the following two days.

4.3 NAIADES platform integration

The integration of the tool into NAIADES platform requires it to fulfill all the NAIADES integration requirements: NAIADES data structure and signature, tokenization, containerization of the tool for installation and APIs and communications protocols amongst components. The last subsection is dedicated to draft a plan for a quick integration.

4.3.1 Common Data Models

All data managed inside NAIADES platform follows the structure defined by NAIADES data models. Any service/tool aiming to use its data must know the NAIADES data models availability. The data models are represented in JSON format so the tools must be able to read and understand the data models; and also to be able to create a correct formatted data model in order to submit new data to the platform.

The Water Quality Forecast Tool will make use of the WeatherObserved, WeatherForecast, WaterQualityObserved and WaterQualityForecast datamodels. The first three data models are basically FIWARE data models[4] with some extra attributes whereas the WaterQualityForecast datamodel is a new datamodel based on the FIWARE WaterQualityObserved datamodel and the WeatherForecast.

[4] https://www.fiware.org/developers/data-models/

4.3.2 Data signature

Data signature is required to guarantee the data has not been manipulated and it allows to analyse all the data movements inside the platform[5].

Data signature is a service provided by one of the NAIADES modules. Each service sharing data with the platform must integrate the Data Signature module (KSI SDK). Once the module is installed the formatted water quality predictions will be signed, thus, the final outcome of the Water Quality Prediction Tool will be a data model formatted and signed water quality forecast. Data signing of sent in data and received data integrity control is provided.

4.3.3 Tokenization

The communication with NAIADES DM goes through an Access Control module that includes FIWARE Wilma and Keyrock blocks[6]. The Water Quality Prediction Tool requires to be identified by these blocks in order to be able to send and receive information. The DM will only share data to NAIADES components that have permission to access it. This identification is guaranteed through tokens. A user and password will be assigned to the tool by the NAIADES platform managers. These credentials are used during the first communication between the tool and the DM to obtain a token that will serve as identifier. This token is meant to be used in every communication (put, patch, get).

4.3.4 Containerization

Previous sections describe three mandatory requirements so the tools are able to interact within the platform. The containerization section is about the optimal packaging of the tool so it is easy installed in the Platform. Containerization allows platform managers to quickly install any tool without worrying about compatibilities and missing libraries[7].

NAIADES consortium agreed to use dockers as standard for packaging all tools. Water Quality Prediction Tool is being developed using Python and it will be dockerized in three steps:

1. Environment definition. To ensure all the required Python libraries are installed.
2. Write the docker file so it creates the environment.
3. Include the python script and the models' files.

[5] Emmadi, N., & Narumanchi, H. (2017, January). Reinforcing Immutability of Permissioned Blockchains with Keyless Signatures' Infrastructure. In *Proceedings of the 18th International Conference on Distributed Computing and Networking* (pp. 1-6).
[6] Pozo, A., Alonso, Á., & Salvachúa, J. (2020). Evaluation of an IoT Application-Scoped Access Control Model over a Publish/Subscribe Architecture Based on FIWARE. *Sensors*, *20*(15), 4341
[7] Mao, Y., Fu, Y., Gu, S., Vhaduri, S., Cheng, L., & Liu, Q. (2020). Resource management schemes for cloud-native platforms with computing containers of docker and kubernetes. *arXiv preprint arXiv:2010.10350*

4.3.5 Data uploading-downloading/Context Manager communication

The data model structure can be used to contain data from different sources, each of these sources will create one JSON file with a unique identifier, this combination of JSON and unique id will represent an entity. The Water Quality Forecast Tool will read ten entities: one for the fountain data, one for the real weather data and eight for the weather estimations.

- The Carouge Fountain WaterQualityObserved. There is only one fountain in this use case, otherwise, each fountain will have their own entity with identifiers such as CarougeFountainWaterQualityObserved1, CarougeFountainWaterQualityObserved2 and so on.
- The Carouge WeatherObserved.
- The Carouge WeatherForecast, which will be provided by NAIADES WeatherForecast service. The Water Quality Prediction Tool will read eight of these entities corresponding to the Morning, Noon, Afternoon and Night of the following two days.

The tool will return just one entity per prediction.

- The Carouge Fountain WaterQualityForecast- Tomorrow Morning.
- The Carouge Fountain WaterQualityForecast- Tomorrow Noon.
- The Carouge Fountain WaterQualityForecast- Tomorrow Afternoon.
- The Carouge Fountain WaterQualityForecast- Tomorrow Night.
- And another four for the day after tomorrow.

The other NAIADES services that need to use these entities, should understand the entity identity nomenclature and the data model structure.

The communication with the data manager will be done using NGSI-v2. It is a RESTful API via HTTP. The Water Quality Prediction Tool uses this API through Python requests following the queries specified in the WIKI provided by the DM partner (https://gitlab.distantaccess.com/naiades/naiades-platform-poc). The entities for this use case will be created by the DM partner (UDGA) but they will be updated by the data providers.

- The Carouge Fountain WaterQualityObserved – by UDGA, from Carouge pilot DCA.
- The Carouge WeatherObserved – by CERTH, from WeatherForecast service.
- The Carouge WeatherForecast – by CERTH, from WeatherForecast service.
- All the WaterQualityForecast entities – by AIMEN, from WaterQualityPrediction service.

Data reading will be done through subscriptions to the real time DM module for the three first entities. This means, the DM will send data to the tool each time a new value is uploaded to the platform. The subscriptions to the historical DM module will be required if the chosen machine learning algorithm requires time series input for inference. Otherwise, the historical DM module will be accessed only for training purposes.

4.3.6 Integration

This section shortly describes the integration planning. The specifications of the infrastructure required to be able to run the Water Quality Prediction Tool will depend on the final machine learning algorithm chosen. The simplest algorithm can be run with CPU computational power and at least 8GB of RAM whereas the RNN solutions will require GPU (similar to NVidia GeForce GTX 1080 or superior) to perform smoothly. The integration is expected to take place in four phases:

1. First model deployment. Once the first model has being created two steps will be taken:
    a. The tool will be validated locally against the developer server. This phase should serve to validate the correct use of the entities, the data signature and the interaction with the DM using ngsi-v2 API.
    b. The tool will be dockerized and sent to SIMAVI for installation. During this first interaction with SIMAVI, all the problems regarding the tool integration in the production platform will be attended.
    c. Tool performance evaluation in the production platform.
2. Second model deployment.
    a. Once the second model, using one of the deep learning tools is ready, the tool will be containerized including the new models and it will be sent to SIMAVI for installation.
    b. Second model performance evaluation in the production platform.
3. Third model deployment. The same process as phase 2 will be followed using the best performance model.
4. Improved deployment. By M30 (November 2021), it is expected to have an improved model using all the data generated during the next year. The same process will be followed to evaluate its performance.

This plan relies on the availability of data on both platforms (development and production servers).

5 Next steps

Currently, the first prototype of the tool is being developed. The required tools to connect to the platform are still under development. This first prototype contains the less computational demanding models created with water quality data freely available online, thus, the objective is to have the tool ready to test all the methods that surrounds the model: data models transformation, data signature, tokenization, DM communications and platform integration. The future steps will require the creation of new models with weather data and different algorithms as well as models upgrades using real data, this is, the model will be retrained periodically using newly collected real data.

- Create the first model (SVR) with real data.
- Use the real data to also create RNN models and compare results.
- Add weather data to the best performance model and integrate it in the platform.
- Periodical model upgrades with new real data until October 2021.

6 Conclusions

This report includes the description, current state and planning of the Water Quality Prediction Tool. This tool is mainly based on machine learning algorithms that uses water quality parameters and weather forecasts to predict future values of specific water quality parameters. Its performance, as all machine learning applications, highly depends on the quality and amount of available data. Currently, the available data is null, it has recently started to be stored for some of the use cases, thus, the quality of the models is not expected to be reasonable until mid-2021.

The tool also requires the integration and implementation of specific functions to be integrated within NAIADES platform, such as data model formatting, data signature, tokenization, ngsi-v2 API and dockerization. The tool will be implemented using Python. The first version will aim to test the specific functions and data sharing with NAIADES DM and afterwards, the tool prediction performance will be upgraded periodically in accordance with real-data availability. It is expected to have a stable version by October 2021 which will be presented in final deliverable D5.10.