Towards computational early warning systems for cyanobacteria blooms in fresh-water swimming waters and drinking water reservoirs

Team composition

Academic team leaders:	Roeland Merks (CWI) , Wolf Mooij (NIOO-KNAW)
Owners case study:	Ghada El Serafy (Deltares), Jan-Willem Blokland (Alten),
Participants:	João Moreira, Lisanne Rens, Joost van Pinxten, Shreya Adyanthaya, Umar Waqas
Advisor:	Joaquin Vanschoren

Abstract

In this report from ICT of Industry 2015, we summarize our work on the problem posed by Deltares, which focuses on early warning system for cyanobacteria in lakes. Cyanobacteria are highly toxic and as such are harmful when present in drinking water reservoirs and recreational water bodies. Therefore, we are interested in accurately and timely predicting cyanobacteria blooms. Based on in situ data in Paterswoldsemeer, we performed some data analysis to get a feeling of time and spatial dependencies between parameters. We also deduced parameters of which measurements were unavailable in this dataset, but are known to play a role in cyanobacteria growth. Using this collection of data, we explored machine learning methods, such as artificial neural networks, to predict cyanobacteria growth. We give recommendations on acquiring extra data, more detailed modeling of cyanobacteria and further exploring machine learning methods. We also recommend considering an ontological approach to early warning systems. Finally, we describe future collaborations and endeavors.

1 Company profile

Deltares is an independent institute for applied research in the field of water, subsurface and infrastructure. Our main focus is on deltas, coastal regions and river basins. Managing these densely populated and vulnerable areas is complex, which is why we work closely with governments, businesses, other research institutes and universities at home and abroad. Our motto is Enabling Delta Life.
Alten Nederland is a leading consultancy company in technical and scientific disciplines, in areas reaching from technical and scientific software development to mechatronics and robotics. Alten is all about technology. Clients are prominent R&D oriented companies or government institutes. We work for these clients on innovative technological development projects with more than 300 highly qualified staff.

2 Description study case

The challenge is to construct an initial predictive, computational model of toxic scum formation of the fresh water cyanobacterium Microcystis aeruginosa, prepare it for the assimilation of satellite data and for integration into existing predictive models for surface water quality. Buoyant cyanobacteria accumulate in dense scums at the lake surface (see picture), forming blooms in which cell densities go up to extreme levels. Because cyanobacteria produce many bioactive compounds, some of which are highly toxic, a minute volume of surface scum material in drinking water reservoirs and recreational water bodies can harm humans and animals. Solving this challenge will enhance robust early warning systems of scum formation and may eventually lead to prevention of scum formation.

The main challenges are (a) how to interface ecological modeling approaches (using, e.g., ordinary and partial differential equations) with existing surface water modeling approaches, (b) how to inform the model using data assimilation techniques. A first model could consist of a simple exponential or logistic growth model in realistic geometries (e.g. obtained from Google maps or from satellite data), based on the availability of nutrients, the amount of sunlight and the temperature. The team could then develop means to obtain those parameters from satellite data, and to assimilate satellite-based observations of blooms into the models. A second phase (only reached depending on the progress in the first phase) would refine the ecological models. This would involve building a multiscale computational model that would take into account buoyance of cyanobacterial colonies and eventually also the molecule scale drive the growth and transition into the buoyant phase.

2.1 Provided data

The team has received in-situ measurements from the period of June 5th 2015 up to October 20th 2015 at different measurement stations in the Paterswoldsemeer, Netherlands. The measurement set contains 3725 raw measurements over a period of 5 and a half months.

The measurements concern Water Insight Spectrometer (WISP) data. WISP handhelds take optical measurements and derive water quality indicators such as:

- the concentrations of chlorophyll-a,

- concentration of phycocyanine,
- CDOM (colored dissolved organic matter),
- suspended matter

Each measurement is annotated with the date and time, the name of the operator, the identification of the WISP handheld used, and the station at which the sample has been taken. The stations are annotated with their longitude and latitude, so that their absolute and relative locations can be used. There is no regularity in the station, or the time at which the measurements have been taken. In certain time intervals, the provided data is sporadic in nature.

3 Revised problem statement

We have revised the problem statement such that it matches the capabilities and interests of the participants. The problem statement has been abstracted to: **Predict cyanobacteria blooms with limited data points**.

Due to the limited amount of time, we have decide to make the main research question: **Is it possible and useful to create or learn models that predict cyanobacteria blooms, specialized for a particular area?**

The challenges towards achieving a good predictive approach are as follows:

1. Identifying the significant parameters that influence blooms
2. Identifying what additional data is needed for accurate predictions
3. Finding means to deduce additional parameters from available information or additional experiments
4. Exploring and assessing prediction models
5. Determine how to integrate data in the model

4 Findings

We identified the types of models that may be used and categorized them as:

- Process based models
- Data driven models
- Knowledge based models

Deltares primarily works with process based models, based on partial differential equations (PDEs) describing the complex water system. With such models, Deltares aims to increase forecasting abilities by applying data assimilation techniques such as Ensemble Kalman Filtering. In the development of process based models, a lot of physical and biological knowledge on the system and its environment are needed. However, complex modeling may not even be needed to acquire an accurate prediction model for a specific location, such as an isolated lake.

We therefore did not aim to understand and describe the underlying biology or physical mechanisms of the system. Furthermore, during the workshop week, it was hard to say anything about suitable data assimilation methods for such process based models without a working model.

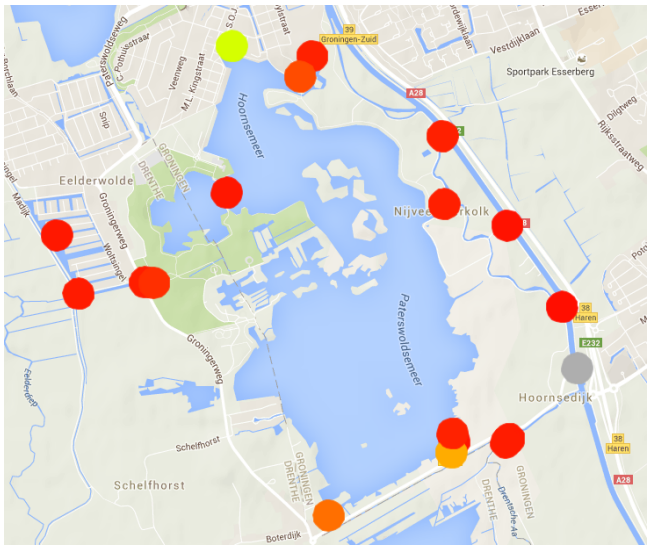


Figure 1: Example of measurement values of phycocyanine concentration (red means bloom, green means no bloom)

Therefore, we decided to focus on data driven models for challenge 4, which now nicely integrates with challenge 5.

4.1 Data reduction, estimation and analysis

The measurements from the WISP data typically contained multiple measurements in a short time interval from the same location. We have considered these as one particular measurement, by aggregating the measurements taking in a single day, at a single station.

The daily weather information contained many aggregated details of the weather during a particular day. We coupled the data of the days in the measurement data of Paterswoldsemeer to the average temperature and wind speed data from the weather information.

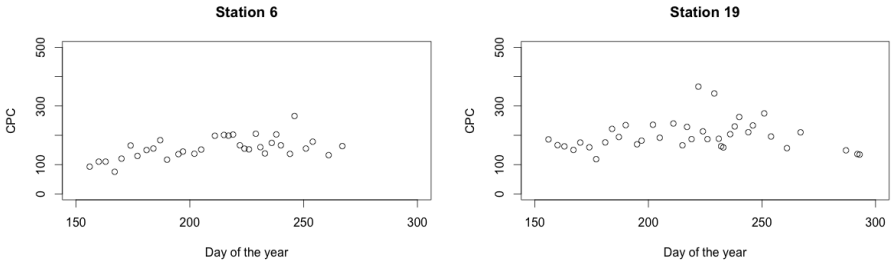
In addition to the provided data, we have used publicly available information such as:

- the satellite and map images from Google Maps, annotated with their longitude and latitude,
- weather information from the Royal Dutch Weather Institute (KNMI), location Eelde, between January 1951 to November 2015.

The information acquired is overlaid on the Google Maps image in Figure 1.

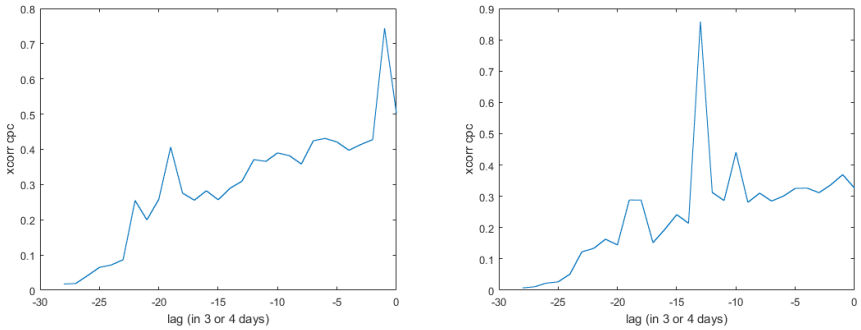
Our strategy to deal with challenges 1 and 2 was to perform data analysis on the available in situ data. This gave us a better idea of the behavior of the system and what kind of information should be integrated with the model. We looked at cross correlations and dependencies of blooms on certain parameters, spatial behavior and temporal behavior and trends. Just as an illustration, we show time plots of CPC (indicating presence of cyanobacteria) at different stations (2), showing slight increases in concentrations late august. Figure 3 shows cross correlation of CPC at different

timepoints. There seems to be a significant correlation of CPC concentrations between timepoints around a month apart. This indicates that maybe the growth of a bloom follows a typical pattern. It would be nice to investigate if this is related to the temperature and light available on these two days. Figure 4 shows the spatial correlations between different stations. This shows that the CPC levels at different stations are highly correlated, indicating that a blooms grow similarly troughout the lake. So, a possible model might not need to have a very dense discretization.



(a) Station 6 (b) Station 19

Figure 2: Time series of CPC at two different stations



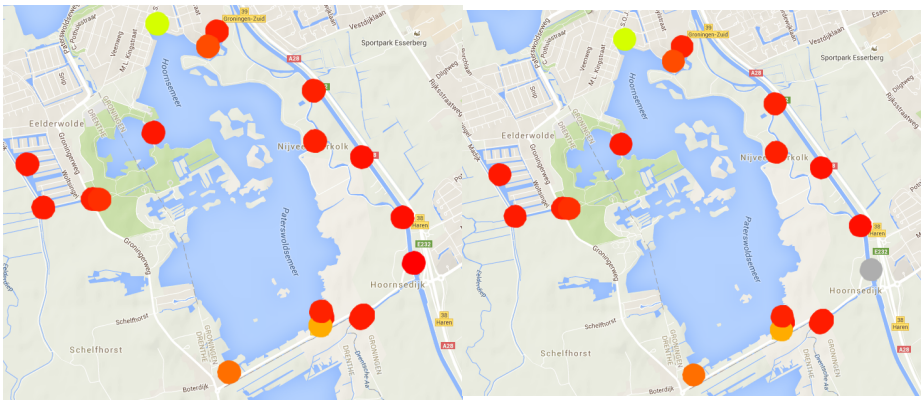
(a) station 6 (b) station 19

Figure 3: Time correlations

4.2 Machine learning

We focused on machine learning techniques, which learn what is the most likely state of the system based on the data. These models do not describe processes or need assumptions and may capture nonlinear dynamics by learning from the data. Methods such as artificial neural networks (ANNs) have been used before [1] to predict algal blooms.

Data driven approaches bring in the benefit of learning the governing principles. Different approaches have different strengths and weaknesses. In this work we explore three approaches, namely, *regression*, *artificial neural networks* and *random forests*. We first present the assumptions



(a) station 6 (b) station 15

Figure 4: Spatial correlations

made during the study of the approaches. Then we present the details of the approaches and results.

Assumptions:

1. The language *R* was used to perform the experiments. All scripts are submitted along the report.
2. The data was normalized to have *mean=0* and *variance=1*. Such a normalization is recommended by most out-of-the-box *R* packages.
3. The provided data was for the *Lake Paterswoldsemeer*. The data was collected from different points called *Stations*. The results shown in this report are from Station 19. The idea is that if we can make 'good' models for a single station later on they can be extended for multiple stations.
4. The data is partially *in-situ* (time, chl, cpc, tsm, latitude, longitude) and partially approximated from Dutch Weather Institute *KNMI* (wind speed, temperature).
5. The learning techniques were used out-of-the-box and were not fine-tuned except explicitly mentioned details. The idea is to first find out how the techniques perform relative to each other and leave the fine tuning as future work. The relative performance is assessed, for example, by comparing mean squared error and sum of squared error.

1) Linear Regression

Linear regression is used to learn linear models for a prediction problem. We used the *lm* method in R to learn a linear model. The call and the summary of the learned model is shown in Figure 5.

As the summary shows, the lm was provided with CPC, $Time$, $Time^2$, $Latitude$, $Longitude$, $Mean\ Temprature$, Chl and *mean windspeed*. The learned model finds out that the longitude and latitude values do not change and are thus not useful for a model for a station.

The polynomial regression reports the *sum of squared error* (SSE) as 3.22.

```
Call:
lm(formula = grouped$CPC ~ grouped$Time + I(grouped$Time^2) +
    grouped$Lat + grouped$Lng + grouped$MT + grouped$Chl + grouped$MWS)

Residuals:
       Min          1Q      Median          3Q         Max
-0.29438669 -0.05633590 -0.00072968  0.06182886  0.48714520

Coefficients: (2 not defined because of singularities)
                    Estimate   Std. Error  t value             Pr(>|t|)
(Intercept)        0.251239662  0.016046813 15.65667 < 0.000000000000000222 ***
grouped$Time      -0.013906234  0.011013060 -1.26270             0.208083
I(grouped$Time^2) -0.014585708  0.008204417 -1.77779             0.076872 .
grouped$Lat                NA           NA       NA                   NA
grouped$Lng                NA           NA       NA                   NA
grouped$MT         0.015122932  0.013361822  1.13180             0.258996
grouped$Chl        1.038697889  0.039695915 26.16637 < 0.000000000000000222 ***
grouped$MWS        0.050953229  0.012430605  4.09901           0.000059094 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 5: The call for linear regression and the summary.

2) Artificial neural networks

Artificial neural networks are a systems of networks, similar to the nervous systems of animals, used to approximate unknown functions depending on large data inputs. They consist of nodes called neurons that are connected with each other by edges having weights associated with them. Starting from random weights or initial weights assigned to the edges, the machine learning process learns the near-accurate weights from the input data.

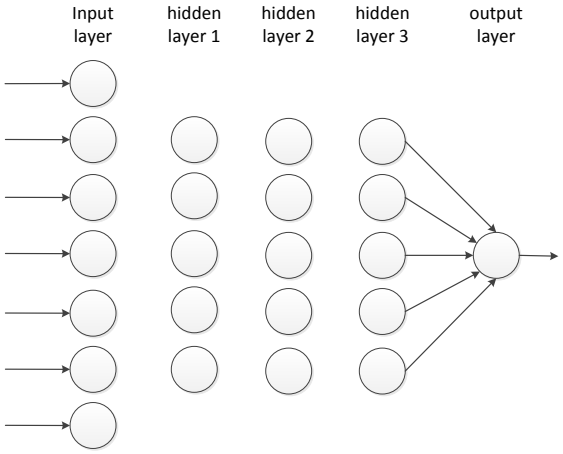


Figure 6: The structure of the artificial neural network used for predictions.

We experimented with different structures of the artificial neural networks. The structure of the network with least error is shown in Figure 6. The network has 3 hidden layers. The variables on which the network was trained are CPC, $Time$, $Time^2$, $Latitude$, $Longitude$, $Mean\ Temprature$, Chl and *mean windspeed*. The output is the predicted CPC value. The *sum of squared error* (SSE) for the network is 6.43. The

network had 1642 training iterations (aka steps).

3) Random forests
Random forests is an ensemble learning technique mostly used for classification and regression. Random forests consist of many decision trees. Decision trees are known to over-fit while learning and forests help in not over-fitting the model by learning trees on different aspects of the data. Then a voting mechanism is used to select a prediction out of the many predictions done by the individual trees. Figure 7 shows an example of a random forest.

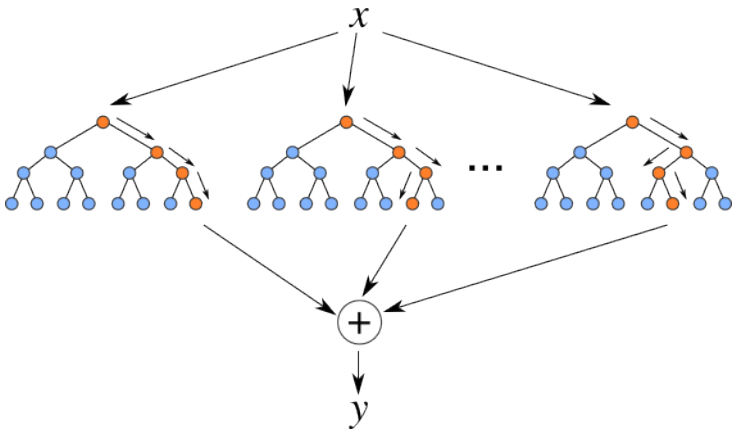


Figure 7: An example of a random forest.

We have used the *randomForest* library of R to learn over the provided data. Figure 8 shows which variables are important for better prediction using two different metrics (IncMSE and IncNodePurity). As the figure shows, time, mean temperature and wind speed are most important factors for prediction. Unfortunately, we did not manage to get error estimates for random forests.

5 Recommendations for company

Per subject the sections below describe the recommendations.

5.1 Machine learning methods

We have investigated three machine learning techniques, namely, *linear regression*, *artificial neural networks* and *random forests*. The language R was used to test out of the box performance of the models under the assumptions described in Section 4.2. Following might be considered by the industrial partner of this workshop.

1. Normalizing the data helps in better performance of the models.
2. The linear regression and artificial neural networks are fast to train and they provide metrics like *sum of squared error* (SSE) as a result of training. Having SSE on hand makes the comparison of the models easier.

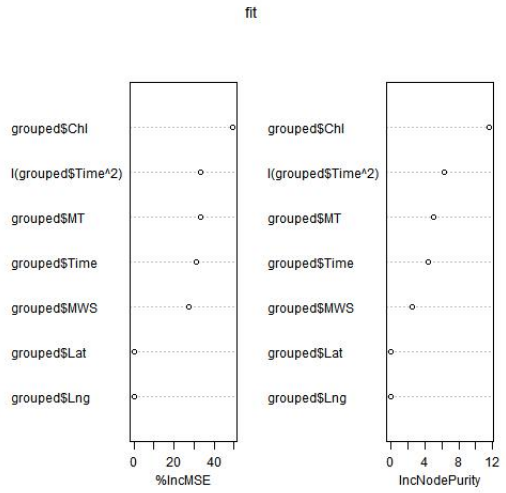


Figure 8: An example of a random forest.

3. Linear regression has better accuracy than artificial neural networks to predict for a single lake.
4. Random forest helps to access which parameters are useful for prediction.

Further investigation is required to fine tune the models and get make them more accurate. Further more, lifting off assumption specified in Section 4.2 is left as future work.

We recommend that Deltares leverages the Crowd-Sourcing platform Kaggle[1], in case they intend to pursue machine learning as a viable alternative to the first-principle models. Kaggle is a community website that allows for competitions on data science for particular usages. Example competition and datasets on Kaggle include; recognizing handwritten digits, finding country development indicators from all kinds of data sources.

5.2 Cyanobacteria growth/transition model

With the right environmental conditions, cyanobacteria can assemble into a multicellular aggregate which then start to float. At this point, the cyanobacteria becomes harmful. So, to predict the growth of cyanobacteria, such a state transition should be better understood and incooperated in the model. This could for instance be done by using the Cellular Potts Model [2], which can describe a group of interacting cells that move through the water.

5.3 Ontology

As described previously, the high-level requirement in this case is a computational early warning system (EWS) for cyanobacteria blooms in freshwater swimming waters and drinking water reservoirs. In disaster man-

[1] www.kaggle.com

Figure 9: Components of an EWS

agement literature, an EWS is defined as: A chain of information communication systems comprising sensor, detection, decision, and broker, in the given order, working in conjunction, forecasting and signalling disturbances adversely affecting the stability of the physical world; and giving sufficient time for the response system to prepare resources and response actions for minimizing the impact on the stability of the physical world [3]. An EWS can be classified as an emergency management system, playing an essential role in disaster risk reduction by supporting emergency agencies in their decision making processes. The common architecture of an integrated EWS is illustrated in Figure 9, for a detailed description refer to [3].

In Deltares case, the focus was the detection component for monitoring and predicting the formation of cyanobacteria blooms. Usually, situation identification techniques are used to support the detection component. Those techniques can be classified in two types: specification-based (top-down) and learning-based (bottom-up) techniques [4]. In specification-based, the patterns that characterize a situation are defined a priori with expert knowledge, typically building a situation model and, then, reasoning on it with input sensor data. The latter technique applies artificial intelligence (e.g. machine learning) to identify situations from the available data, such as linear regression, artificial neural networks and random forests. Learning-based techniques were best suited for the case proposed by Deltares. However, a hybrid approach is suggested whether a domain expert is available [4], in our case a biologist or an algae expert. That way, represented and derived knowledge from the available data sources can complete each other. Therefore, some recommendations can be made towards an hybrid mechanism for the detection of toxic scum formation of cyanobacterium Microcystis aeruginosa. In the recent years numerous ontological approaches have gained attention as specification-based techniques for EWS. Ontologies can formally explicit the domain knowledge through formal axioms and constraints by supporting primitives of modelling constructs. Reasoning can be performed to derive new knowledge due to the formalism of ontologies. For some examples of the use of ontologies in deriving new facts refer to section 4.3 of [4]. Therefore, we suggest the creation of an ontology to explicit the knowledge regarding the formation of cyanobacteria bloom and how data is gathered. A high-level draft of this ontology is illustrated in Figure 10. On top, we designed stations monitoring chlorophille and CPC measures through sensors. On bottom, we described that a lake can have individuals of cyanobacterium Microcystis aeruginosa composing cyanobacteria blooms. Other specific concepts must be added in this ontology, as well as rules to define both structural and behavioral aspects. Domain experts must be consulted.

In particular, we suggest an approach for an holistic EWS for cyanobacteria blooms by applying a framework that is being developed under a PhD project, described in [5]. This framework stresses the notion of situation and it is well-founded on a research (of more than a decade, including PhD and MSc thesis). It provides the components to the creation of EWS with a hybrid situation identification mechanism for the detection of disaster

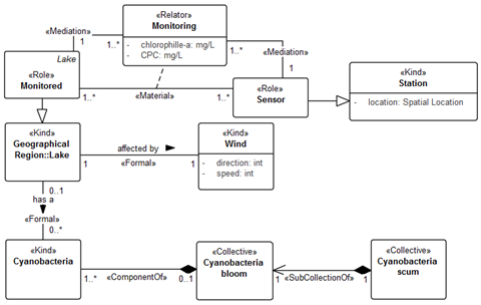


Figure 10: A draft of the ontology for cyanobacteria bloom monitoring

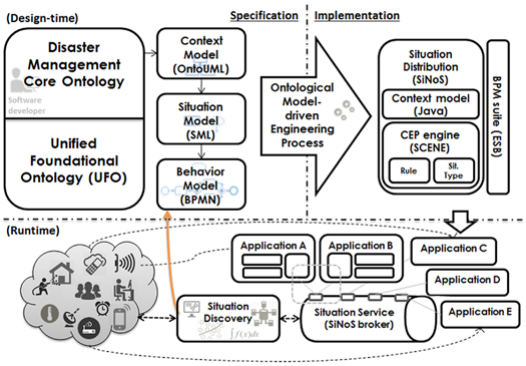


Figure 11: Framework to support the development of EWS through an hybrid situation identification

situations. Figure 11 depicts these components through a software engineering point of view, where the top represents the design time of the EWS, while the bottom represents the runtime.

As usual in systems engineering, the design-time considers the interaction of specification and implementation phases supported by a software engineering methodology. Specification relies on modelling languages to explicit structural (e.g. UML) and behavioral (e.g. BPMN) aspects. On top of them, meta-models must formalize the modeling constructs and their admissible relations. In our framework the Unified Foundational Ontology (UFO) was chosen as this meta-model. UFO provides and ontological language called OntoUML for structural modeling. It is used to describe the main emergency management structural aspects through the disaster management core ontology (coined OntoEmerge). The software developer and domain experts can extend OntoEmerge to more specific concepts, such as cyanobacteria blooms and scums. Then, they can specify the rules (patterns) that represent the situations to be detected with the Situation Modeling Language (SML). For each situation of interest (called situation type) a reaction can be also specified with BPMN. Bellow (Figure 12) we illustrate, on left, the rules about chlorophille concentration in water, and, on right, the actions (procedures) to be performed for each situation type: for red alert evacuate the lake, for yellow alert contact the authorities and for green alert keep routine operation.

The framework is developed under the model-driven engineering paradigm,

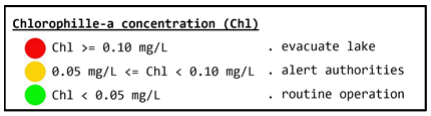


Figure 12: Emergency situations for chlorophille concentration in water and their procedures

where the implementation code is (semi) automatically generated by the models. A rule-based engine for complex events processing, coined Scene (based on Drools), supports situation detection. A distribution mechanism, coined SiNoS, supports the sharing of information about the detected situations. A Business Process Management (BPM) suite supports the execution of the situations reactions. In runtime, SiNoS acts as a publish-subscribe middleware following service-orientation architecture (SOA). At last, but not less important, we can fit the existing learning-based approach for cyanobacteria blooms formation in the situation discovery component. The correlations and predictions resulted by the learning-based part can be used as feedback for domain experts and software maintainers, playing an essential role towards a self-adaptive EWS. The main benefits on applying this framework to build an holistic EWS of cyanobacteria blooms formation can be summarized as:

- Explicit knowledge from domain experts with an ontological approach, supporting common understanding with graphical modelling, especially for non-IT experts.
- The EWS can be self-adaptive because the rules can be changed in runtime due to the rule-based engine nature.
- It leverages distributed processing of big data by using infrastructure in the cloud.

6 Follow up/outlook

The companies want to use these workshop result to submit a project proposal at STW. Also, they want to write a review/state-of-the art paper on methods to predicts cyanobacteria where they explore the possibilities posed in this workshop. Also, machine learning methods could be tested on 'fake' data acquired from existing models describing algae growth in lakes. This might give an idea of the accuracy of such methods and which and how much data they need as an input.

Deltares wants to include cyanobacteria population dynamics in their water quality model. They are setting up a student project where such a growth model can be compared to the existing model, where no population dynamics is considered. Data for three different lakes have been assembled in 2014 en 2015, which can then be used to validate such models.

References

[1] Hugh Wilson and Friedrich Recknagel. Towards a generic artificial neural network model for dynamic predictions of algal abundance in freshwater lakes. *Ecological Modelling*, 146(13):69 – 84, 2001.

[2] F. Graner and J.A. Glazier. Simulation of biological cell sorting using a two-dimensional extended Potts model. *Phys. Rev. Lett.*, 69(13):2013–2016, September 1992.

[3] N Waidyanatha. Towards a typology of integrated functional early warning systems. *International Journal of Critical Infrastructures*, 6:31–51, 2010.

[4] J. Ye, S. Dobson, and S. McKeever. Situation identification techniques in pervasive computing: A review. *Pervasive and Mobile Computing*, 8:36–66, 2011.

[5] J. L. R. Moreira, L. Ferreira, M. V. Sinderen, and P. D. Costa. Towards ontology-driven situation-aware disaster management. *Applied Ontology*, 10:339–353, 2015.