Quest Journals
Journal of Research in Environmental and Earth Science
Volume 5~ Issue 2 (2019) pp: 01-09
ISSN(Online) : 2348-2532
www.questjournals.org

Research Paper

Turbidity forecasting in the Delaware River

Pere López Brosa, [1,2] Antonio Monleón-Getino, [1,2]
Javier Méndez Viera[1,3], *Francisco Lucena Gutiérrez*[1,3]
[1](*BIOST3 Group. Section of Statistics. Department of Genetics, Microbiology and Statistics. University of Barcelona)*
[2](*Section of Statistics. Department of Genetics, Microbiology and Statistics. University of Barcelona)*
[3](*Section of Microbiology. Department of Genetics, Microbiology and Statistics. University of Barcelona)*
[4](*GRBIO. Research Group in Biostatistics and Bioinformatics)*
Corresponding Author: Antonio Monleón-Getino

ABSTRACT:*Turbidity in water sources is a concern towater treatment plants. Aregression-based methodology for prediction of turbidity in rivers is proposed and an application on data from Delaware River is shown.*
KEYWORDS:*Turbidity forecasting, Rivers, Water sources.*

Received 17Aug., 2019; Accepted 31 Aug , 2019
Published with open access at www.questjournals.org

I. BACKGROUND AND OBJECTIVES

High turbidity in source water is a problem for purification and disinfection of drinking water, leading to higher costs and often-temporaryturning off treatment plants (Mather 2016). Therefore, a reliable turbidity forecasting can improve water management efficiency. Furthermore, in a changing climate context, relating precipitation patterns to water streams turbidity may allow to use projected trends for precipitation to predict future negative impacts of turbidity on water treatment plants availability.

The purpose of this study is to establish a turbidity forecasting methodology for water streams using precipitation as a predictor.

1) Delaware River

Data from the Delaware River wasused to demonstrate the methodology. Delaware River is a Middle Atlantic major river ofthe United States, flowing from Catskill Mountains in the state of New York, through Philadelphia and New Jersey into Delaware Bay in the Atlantic Ocean. Trenton, New Jersey, isthe point where the river meets tidewater and divides the course of the river between upper and lower Delaware. The basin of the Delaware River at Trenton is about 17600 km^2 and its average discharge 337.6 m^3/s (years 1913 to 2016). The main stem of the river is un-dammed although there are two dams in the two branches that form its headwaters.

Turbidity at Trenton is usually low with just moderate peaks. Figure 1 shows a histogram of turbidity at Trenton (in Nephelometric Turbidity Units) for the period 2007-2017.

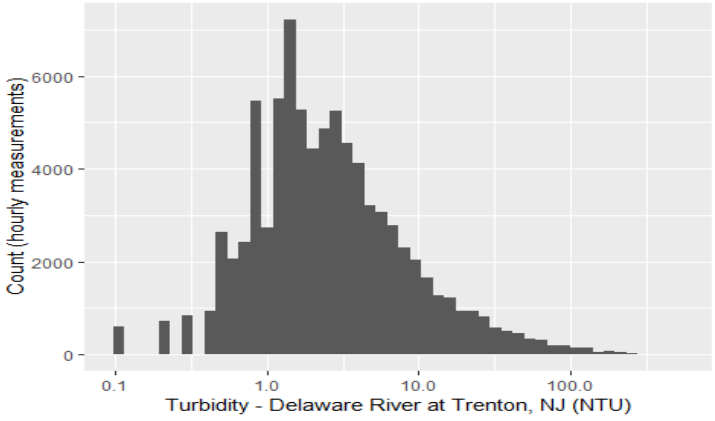

Figure 1: Turbidity at Trenton (in Nephelometric Turbidity Units) for the period 2007-2017.

III. METHODOLOGY

Linear Regression is a powerful statistical technique and can be used to understanding factors influencing turbidity and to evaluate trends and make estimates or forecasts (See Monleón-Getino and Rodriguez (2017) and Monleón-Getino and Canela (2017)). Linear regression refers to a model that can show relationship between two or more variables and how one can impact the other. Basically,regression tries to explain how the variation in the "dependent variable" (Y) can be captured by change in the "independent variables" (X_i).

Another possibility is to use non-linear regression, but with the consequence of using much more complex and artificial models; previous experiences have ruled out their use at this moment (Mendez et al, 2018).

For the construction and analysis of the models, the statistical package R, version 3.5 has been used (R Core, 2019)

One linear model was fitted for each forecast horizon to obtain forecasted turbidity and local regression was performed on residuals to get confidence intervals. Since response variable (turbidity) was highly skewed, a logarithm transformation has been appliedbefore regression analysis. Logarithm transformation was also consistent with results from Box-Cox analysisand decimal logarithms were chosen to ease interpretability.

Predictors were:

- Recorded turbidity for previous eighthours.
- Recorded hourly precipitation for previous fourdays (96 hours).
    - Initially, fourdays before the moment the forecast is made were used.
    - Alternatively, a simulated perfect weather forecast was used to predict turbidity, for simulation 4 days of precipitation records previous to the forecasted moment were used.
- To account for seasonal and daily effects, harmonic functions (sine and cosine) with period oneyear to a sixthyears and oneday to half day were used.

1) Box-Cox analysis

Box-Cox method has been applied to assess better transformation of response (turbidity) using package MAAS (Venables 2002). Since we use a different linear model for each different forecast horizon, lambda for maximum likelihood has been determined for each one (Figure 2).

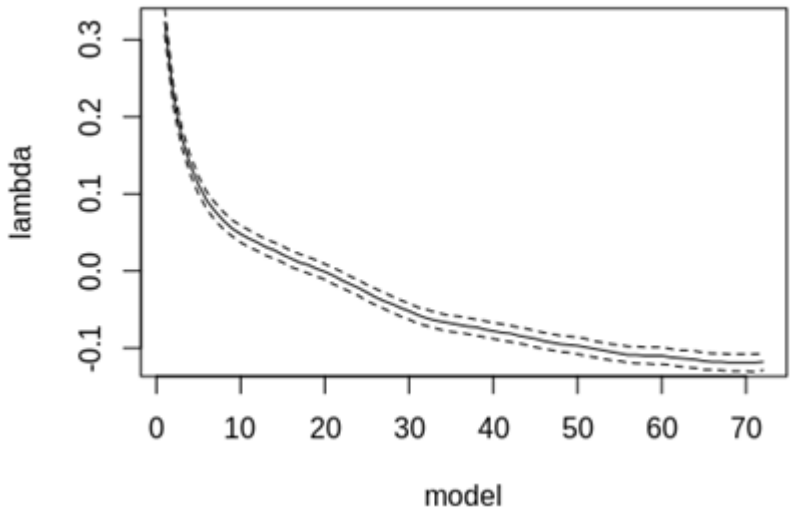

Figure 2: Box-Cox lambda with maximum likelihood (with 95% confidence intervals) vs forecast horizon (in hours).

The analysis suggests different optimal transforms for different forecast horizons: cubic root for short horizons and logarithm (or beyond) for longer ones. However, to keep the methods consistent and results explainable we chose to use decimal logarithm for all forecast horizons.

IV. AVAILABLE DATA

The USGS's National Water Information System provides a fairly continuous time series of turbidity of Delaware River at Trenton, which was measured hourly from 2007 to present(USGS, 2018). Turbidity measures at Frenchtown (about 50 km upriver from Trenton) are also available, but they have not been used here because measurements only started in 2014.

For the period 2007-2014, 17 weather stations are listed as available at the NOAA's National Climatic Data Center(NOAA, 2017). However, hourly precipitation time series for these stations have large gaps at best and are nearly completely missed at worst. Furthermore, since data availability for different stations happens at different periods, for any sizeable given sets of time points and stations, the number of complete cases is small. In our experience, this challenging scarcity of complete observations appears to becommon across meteorological sources.

We have selected as predictors the fiveweather stations in the basin with most observations.

Figure 3 plots availability of precipitation data of weather stations in upper Delaware basin and turbidity measures at Trenton for each date, highlighting the five stations used in our models and overlap of availability of those five stations.In figure 4 geographical location of weather stations is shown on a map of the basin.

Since linear regression and other common methods only allow complete cases to be used, when selecting the weather stations, we faced a trade-off between number of predictors (stations) and number of complete cases. In the end, less than 20% of recorded data remained available for the model, potentially impairingits fitting.

Figure 3: Plot of availability of precipitation data of weather stations in upper Delaware basin and turbidity measures at Trenton is shown. Complete cases for the whole set of the fivestations with more observations are highlighted, as those are the cases used to fit the regression model.

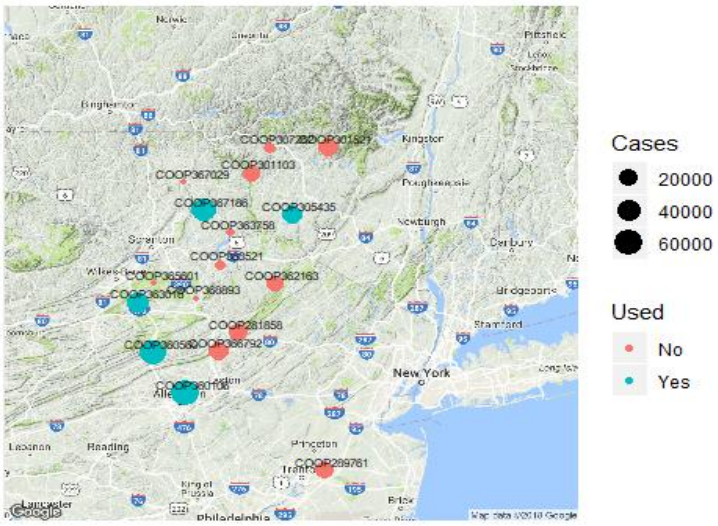

Figure 4: Location of weather stations. It is shown the number of cases recorded for each station and whether data from that station has been used as predictor.

V. CONFIDENCE INTERVALS

Performing regression analysis with hundreds of predictors and tens of thousands of cases implies huge load times and huge memory requirements to be addressed by usual methods (as the common *lm* function in R), thus memory saving methods may be needed (like *biglm*). One drawback of such methods is the lack of confidence intervals for the prediction in the built-in predict function because of computation of such intervals are at least as memory-hungry as model fitting.

As an efficient way to compute prediction confidence intervals, we have used locally weighted scatterplot smoothing (LOESS) of residuals on predicted values. By using the function *loess.sd* from R package *msir* we compute the loess smooth for the mean error plus and minus 1.96 times the standard deviation function to get 95% confidence interval on single errors (Scrucca, 2011); thereby, confidence intervals for turbidity single predicted values wereobtainedby adding the lower and upper bound of confidence intervals for residuesto the predicted values.In figure 5 upper and lower limits of confidence intervals of residues are plotted against predicted value, for four different forecast horizons.

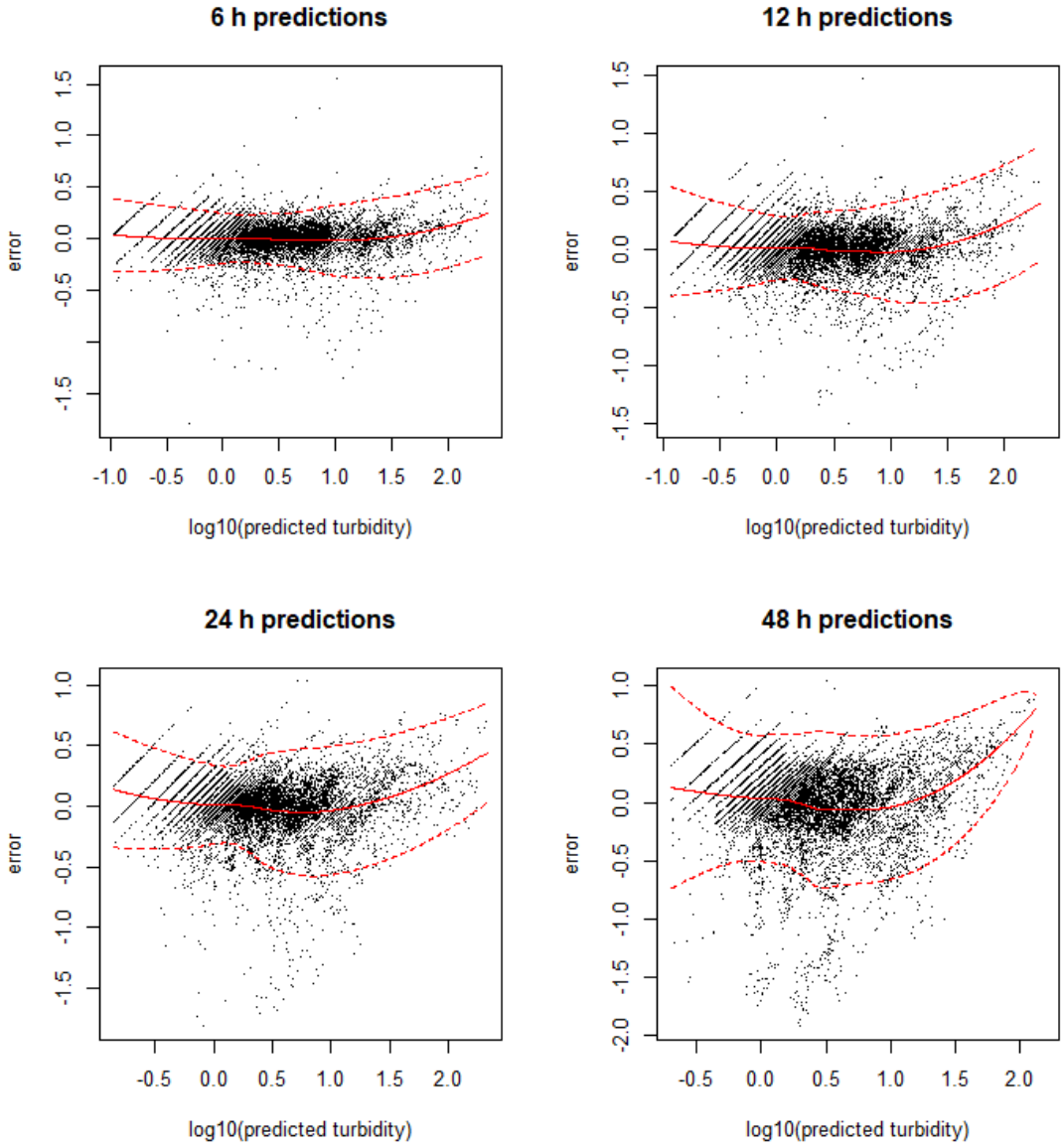

Figure 5: Predicted values and 95% confidence intervals on residuals.

VI. EXAMPLES

A couple of high turbidity episodes are used to display the approach. For each one the following graphics are shown:

- Graphics and maps of precipitation recorded in each weather station.
- Forecasted turbidity evolution from several moments in the episode, using and not using simulated precipitation forecasting. Confidence intervals on each forecast are provided.

1) March 5th to 11th 2011

In figures 6, 7 and 8 are shown two turbidity episodes that took place overfew hours after two rain episodes in the basin. Once precipitation had been registered, the rise of turbidity was predicted correctly by the model. When a perfect rainfall forecast was simulated, the forecast kept its accuracy for an extended period (figure 9).

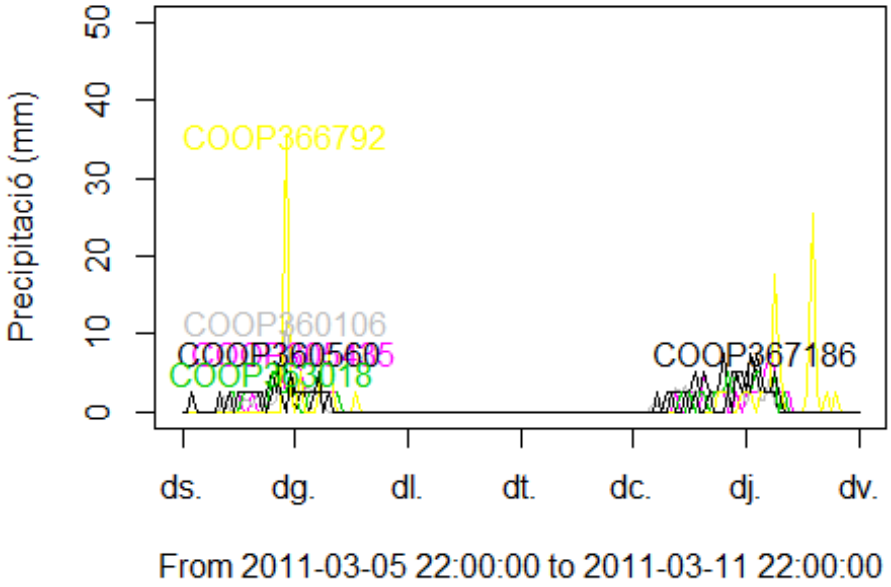

Figure 6: Precipitation (mm) during the episode.

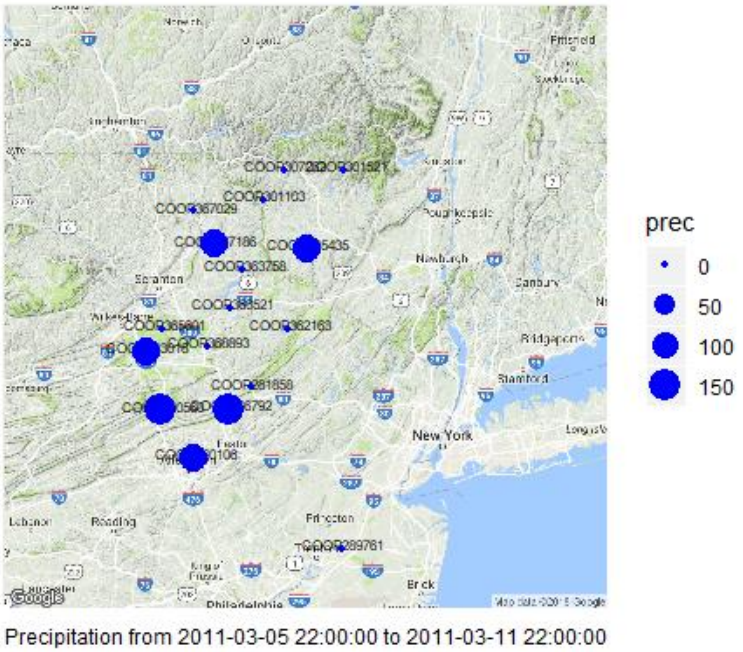

Figure 7: Precipitation (mm) during the episode.

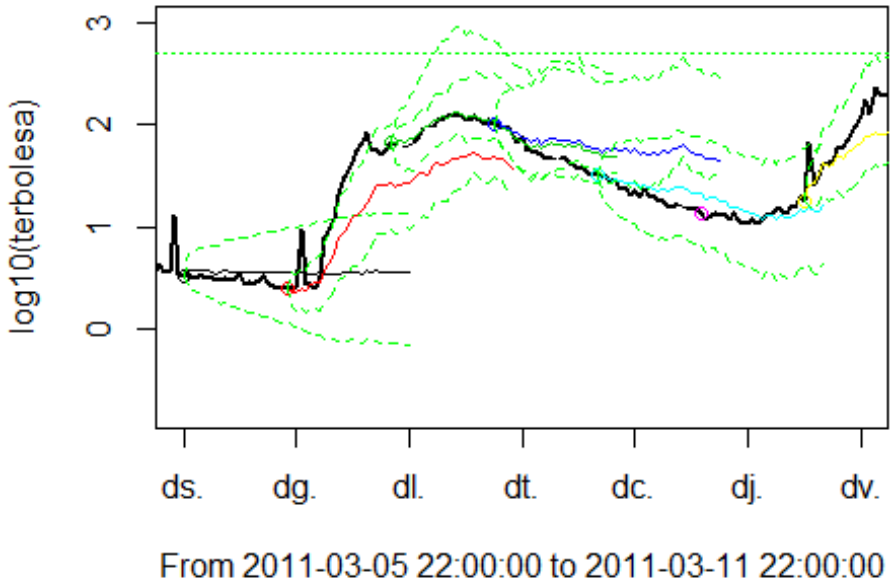

Figure 8: Measured turbidity (black thick line) and forecasted turbidity (colour lines) with 95% confidence intervals. Forecasts are up to 48 hours, using previous hours of registered precipitation as predictors (no weather forecast used nor simulated).

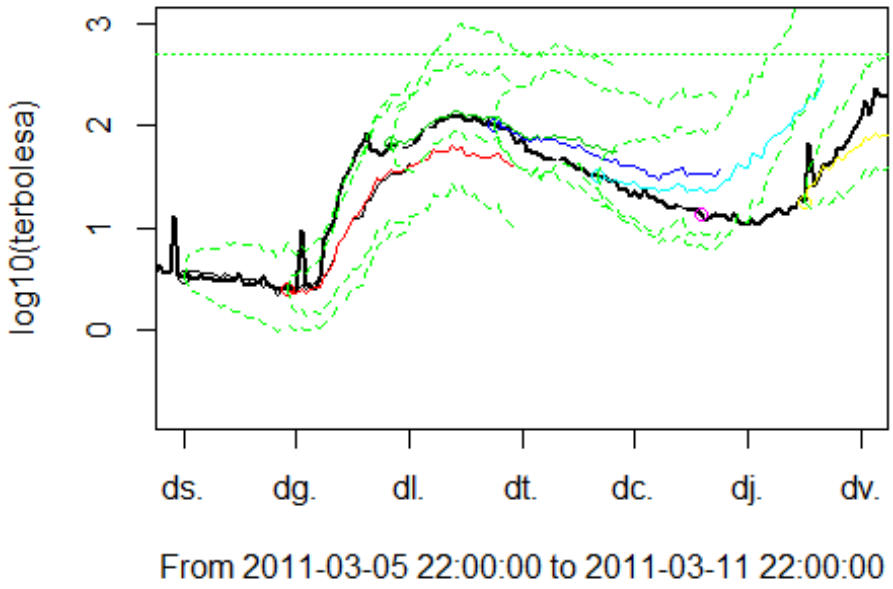

Figure 9: Measured turbidity (black thick line) and forecasted turbidity (colour lines) with 95% confidence intervals. Forecasts are up to 48 hours, using previous hours of registered precipitationandfuture registered precipitation (as a simulated perfect rainfall forecast) as predictors.

2) May 6th to 9th 2009

Turbidity rose after one rain episode (figures 10 and 11). Models predicted a turbidity rise but severely underestimated its value (figure 12) - although models with a simulated rainfall forecasting performed a little better (figure 13). When precipitation registers were examined, the heaviest rainfall was found to have happened on a station in the head of the basin that had not been used as predictor in the models. This example highlights the importance of developing methods to use all available predictors even when a large amount of data is missing.

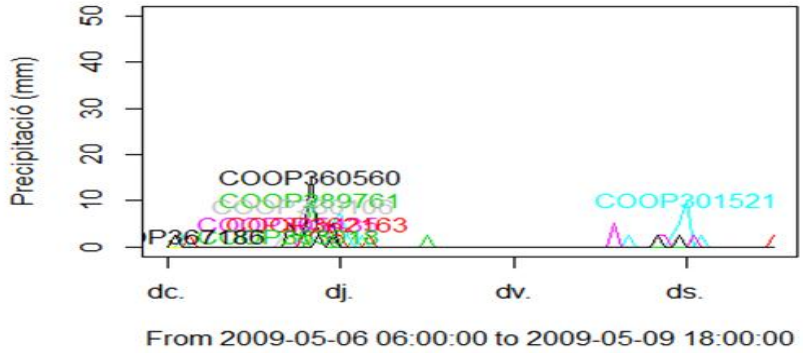

Figure 10: Precipitation (mm) during the episode.

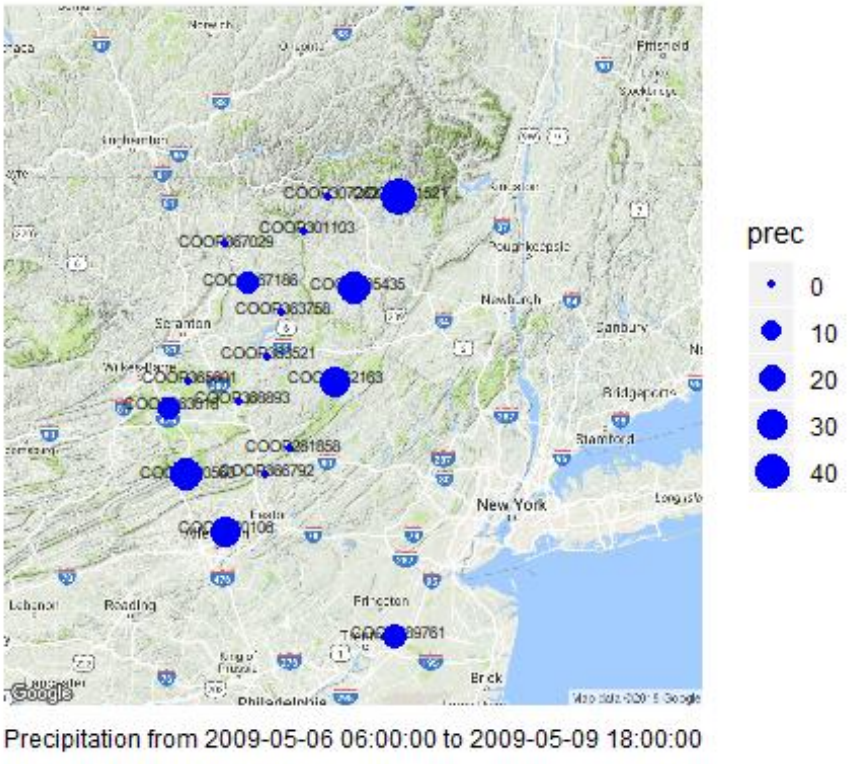

Figure 11: Precipitation (mm) during the episode.

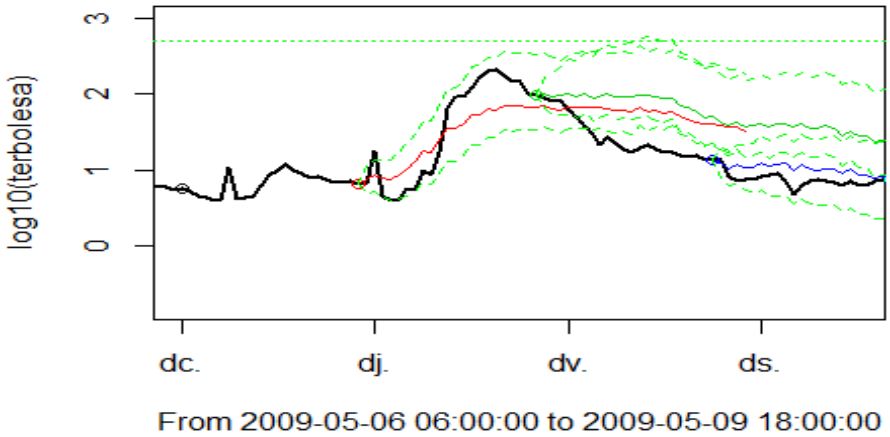

Figure 12: Measured turbidity (black thick line) and forecasted turbidity (colour lines) with 95% confidence intervals. Forecasts are up to 48 hours, using previous hours of registered precipitation as predictors (no weather forecast used nor simulated).

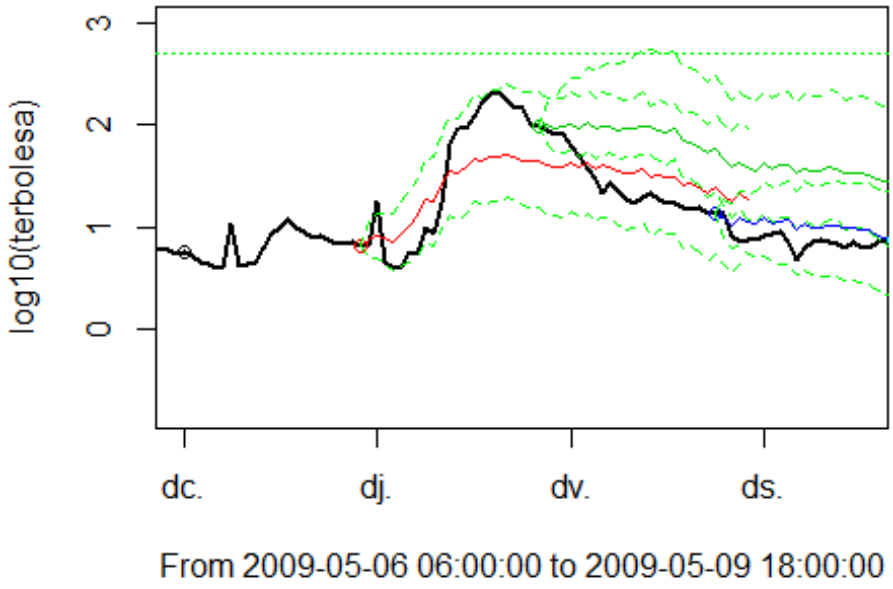

Figure 13: Measured turbidity (black thick line) and forecasted turbidity (colour lines) with 95% confidence intervals. Forecasts are up to 48 hours, using previous hours of registered precipitationandfuture registered precipitation (as a simulated perfect rainfall forecast) as predictors.

VII. CONCLUSIONS

Regression models with the used predictors yielded predictions accurate enough to be used of early warning of turbidity episodes. Moreover, turbidity predictions span enough time to be used by water treatment plants to estimate the expected time of inactivity. Improved models are expected to improve accuracy.

When simulated perfect weather forecasting was used, turbidity estimation gets accurate for extended periods, suggesting that weather forecasts should be used as predictors.

Since only complete cases can be used withlinear regression and other common methods, less than 20% of recorded data remain available for the model, potentially hindering its fitting. Furthermore, episodes of stronger precipitation in areas covered only by stations not included in the model causes a noticeable lack of accuracy of predictions. Further work should be addressed to combine different models to make the best possible use of all observations.

1) Future work

- Selection of variables:
    - Improve trade-off between number of weather stations and number of available observations.
    - Optimize time span of previous measures of precipitation and turbidity to take in account.
- Collection and usage of meteorological forecasts as predictors.
- Application the model to other streams.
- Comparison of methods to produce confidence intervals.
- Use a non linear models or other methodologies like machine learning algorithms.

VIII. CONCLUSIONS

This work was funded by Spanish Ministerio de Economía y Competitividad MEDSOUL478 project (CGL2014-59977-C3-1-R) and the Catalan government (2017 SGR 170).

REFERENCES

[1]. Amanda L. Mather & Richard L. Johnson, 2016. "Forecasting Turbidity during Streamflow Events for Two Mid-Atlantic U.S. Streams," Water Resources Management: An International Journal, Published for the European Water Resources Association (EWRA), Springer;European Water Resources Association (EWRA), vol. 30(13), pages 4899-4912, October.
[2]. Scrucca, L. (2011) Model-based {SIR} for dimension reduction. Computational Statistics & Data Analysis, 5(11), 3010-3026.
[3]. USGS (2018). National Water System Information. USGS 01463500 Delaware River at Trenton NJ
[4]. NOAA (2017). Climate Data Online
[5]. Monleon-Getino, J Canela Soler. 2017. Causality in Medicine and Its Relationship with the Role of Statistics. Biomedical Statistics and Informatics 2 (2), 61-68

[6]. Monleón-Getino T, Casado CR. 2017. Probability and statistics for Science II. EdicionsUniversitat Barcelona
[7]. R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.URL http://www.R-project.org/.
[8]. Mendez J, Monleon-Getino A, Jofre J, Lucena F.2017.Use of non-linear mixed-effects modelling and regression analysis to predict the numberof somatic coliphages by plaque enumeration after 3 hours of incubation. Journal of water and health 15 (5), 706-717
[9]. Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

Antonio Monleón-Getino" Turbidity forecasting in the Delaware River" Quest Journals Journal of Research in Environmental and Earth Science, vol. 05, no. 02, 2019, pp. 01-09