DIGITAL EARTH

HelmholtzZentrum münchen
Deutsches Forschungszentrum für Gesundheit und Umwelt

September 2020
#1

Newsletter of the *Digital Earth* Project

Contributions of HMGU Helmholtz Zentrum München

This newsletter presents the projects of the institutes of the HMGU being involved in activities within the Show Cases or Work Packages of Digital Earth.

Point to Space workflow and application in show case methane

Mahyar Valizadeh & Wolfgang zu Castell
HelmholtzZentrum München - German Research Center for Environmental Health (HMGU)

The fundamental concept of point to space problem is mainly to find a solution to conform multiple data sources and building a map based on them. Data sources at scattered sites (in-situ measurements), gridded data (satellite data, output of numerical models) are combined to predict higher resolution map or at ungauged sites (interpolation, "downscaling") or to simulate kind of "artificial data" for unobserved scenarios.
At the Digital Earth 2nd Annual Meeting a basic rough idea of the point to space workflow based on ML algorithms has been presented and the application of it in showcase methane has been investigated. The idea is to build a workflow that can input multiple covariates from point or gridded datasets and find an estimator for the outcome variable.

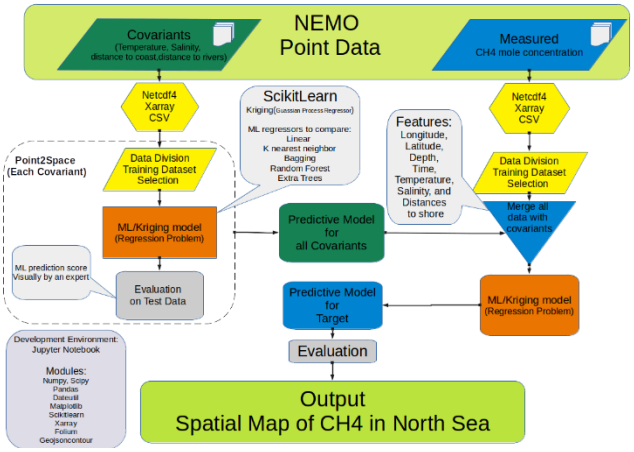

Figure 1: Flowchart for the point to space problem example using the NEMO oceanographic data.

Shown in Figure 1 is the basic workflow and tools used. Figure 2 shows the actual measured methane mole concentration measured overall by multiple sources over several months. A subset of the problem is another point to space problem for the covariates, since we might not have the data for the higher resoluted grid of interest. This has been illustrated in Figure 3 for sea water salinity, sea water temperature, distances of measuring devices to coasts and rivers. For the case of salinity and temperature data is in point data format and for distances is in map format but different resolution than target. Now a prediction ML regression model can be trained and used to predict for the target variable, the methane mole concentration, based on the percentage of the data used as training.

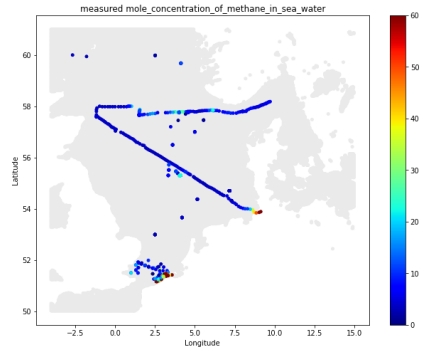

Figure 2: Point measurements of CH4 mole concentraion.

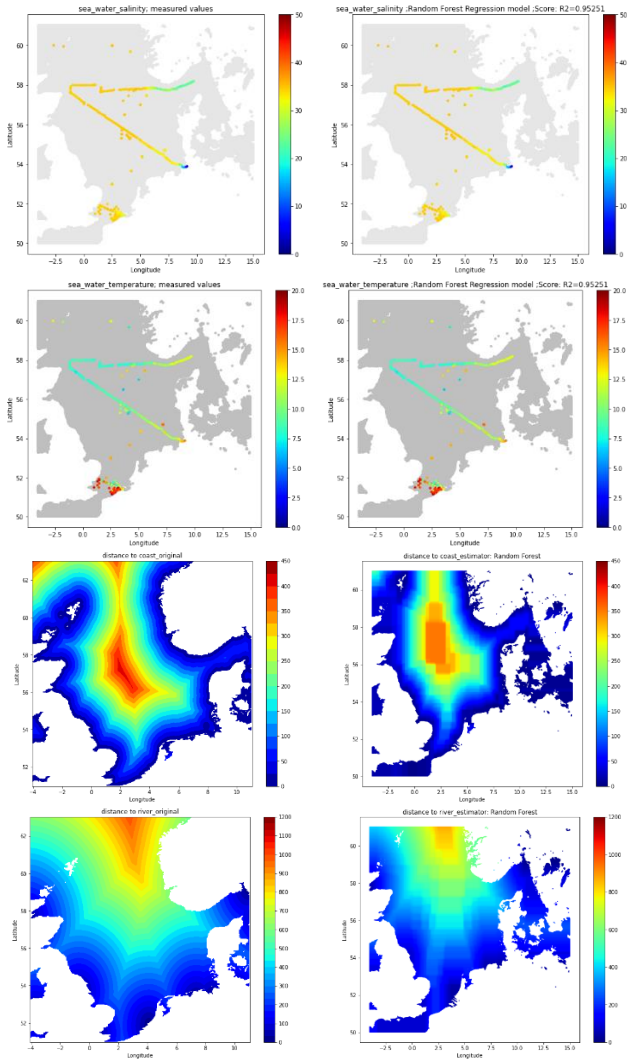

Figure 3: Covariates and their estimate using Random Forest regressor (From top to bottom: Salinity, Temperature, Distance to coasts, and distance to rivers).

For illustration purposes the prediction results of Random Forest regressor on forty percentage used as training data is visualized in figure 4.

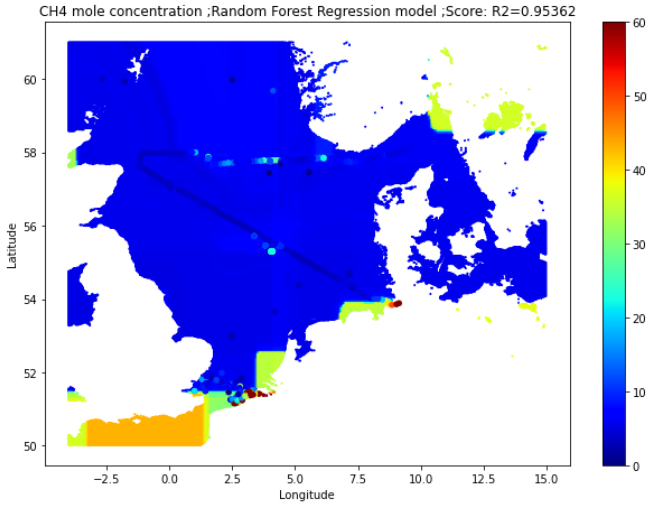

Figure 4: Result of ML predictive model (Random Forest for methane concentration in North Sea.

However, any regressor can be easily implemented and used here, even a self-modified kriging method. At the 2nd annual meeting a comparison of methods Linear, K-nearest neighbours, Bagging, Extra Trees, Random Forest and simple ordinary kriging has been evaluated and it has been shown that the tree methods are very trivial and also has advantage in prediction score for this problem.
To apply the same workflow to other similar problems following considerations need to be contemplated.
Generally, the major issue is to pre-process and clean data since the sources comes from different databases. However, this problem is getting simpler to solve with standardization and metadata sharing. Another issue is that given the nature of the problem the respective expert needs to be consulted specially in selection of covariates.
As with all machine learning regression problem solving it with a regressor can have the curse of overfitting and some mitigation to this issue is to use methods which are taking this issue in consideration.
Normalization is another factor in play since the covariates can be of different order and hence error propagation can be relevant to their scale. A proper way to normalize is also very hard to decide since it can have direct effect on the solution.
In the end, currently there are several issues with this implemented method and a progression of the method will try to solve them. To mention a few of them, building covariance function for cokriging is not easy since the methods with dependant covariates are so sensitive to the input. Again, the evaluation of results using expert knowledge requires deep understanding of the problem since data can be visually appealing but are extremely erroneous.
Uncertainty in data and methods should be evaluated to ensure the final step that the method is in a proper stable solution. This project enjoys the successful new collaborative opportunity between multiple centres involved in Digital earth since the point is to bring knowledge from respectively conforming projects under the same hood.

Analysing flood events with QtAC

Hannah Schrenk[1], Wolfgang zu Castell [1]& Stefan Lüdtke [2],
1 HelmholtzZentrum München - German Research Center for Environmental Health (HMGU)
2 German Research Centre for Geosciences Helmholtz Centre Potsdam (GFZ)

The development of general complex systems was heuristically described in form of the adaptive cycle metaphor (Gunderson and Holling, 2002). According to the metaphor, a system alternately runs through phases of exploitation and conservation and phases of release and reorganization. The development is captured via three systemic variables, potential, connectedness, and resilience. While the connectedness of a complex system increases during the phases of exploitation and conservation, resilience decreases. The system breakdown is accompanied by a sudden and strong decrease in connectedness and an increase in resilience. Figure 5 shows the typical "lying-eight" visualization of the adaptive cycle.

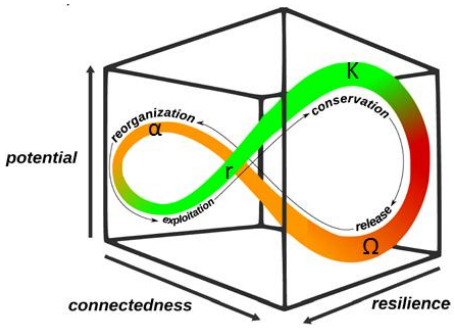

Figure 5: Visualization of the adaptive cycle metaphor.

We developed a method to quantify the adaptive cycle metaphor. The method only requires time series of the system's components' abundance, which serve as a basis to estimate networks of information transfer (Schreiber, 2000). Potential, connectedness, and resilience are then computed as properties of the information networks. This procedure results in a time series of the three systemic variables, giving insights into the system's development during the observation period and possible hints on the future development. By means of the information networks, we can explore the role of individual components in the system and identify drivers of change. The R package QtAC, which is provided in the Digital Earth Gitlab group in a work version, enables a straightforward application of the method.

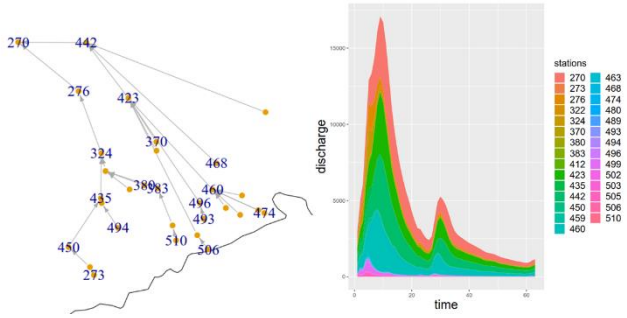

Figure 6: Geographical network of the gauging stations and their discharge of water during the analysis period.

Applying our method to data of the Digital Earth showcase "Flood" allows us to explore flood events from a systemic point of view. We consider a system of gauging station in Saxony, as being represented in Figure 6. The daily discharge values from May to July 2013 serve as the components' abundance data. Figure 2 shows the extent of the flood event that has happened in the river system during this period.
Our results indicate that the river system goes through several phase changes during the observation period. Phases of highly linked information networks like at time point 21, reflected by high connectedness and potential, alternate with phases of loosely connected networks like at time point 24 (see Figure 7).

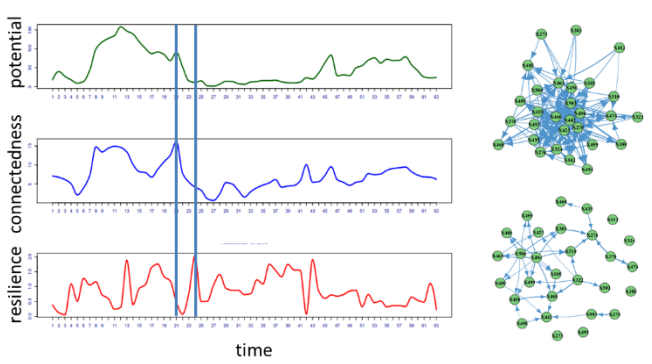

Figure 7: Potential, connectedness, and resilience of the system of gauging stations during the analysis period. Information networks at time points 21 and 24.

In this period, connectedness and potential are comparably low, indicating that the system runs through a phase of release or reorganization. Both during times of high connectedness and times of low connectedness, the information networks reflect the geographical network, more precisely, there is a lot of information transfer from the upriver stations to the ones being situated more downstream. However, during times of low connectedness, the structure of the network changes multiple times. Exemplarily, while at time point 60, stations 276 and 442 take central positions in the network, at time point 61, only station 276 takes a central position (see Figure 8).

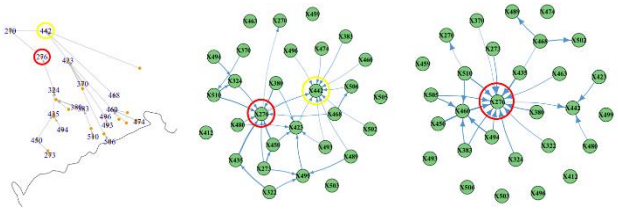

Figure 8: Information networks at time points 60 and 61 in comparison to the geographical network.

This instability is typical for phases and release and reorganization. We assume that with the flood "moving" through the river system, established information transfers can be interrupted, thereby decreasing the number of edges in the information network, and the role of the individual components can dynamically change.
At this point, we are convinced that our method captures system changes being related to the flood event. However, further refinements concerning the choice of parameters will be necessary to deepen the analysis and to receive significant results. We will focus on possible indicators for upcoming flood events as well as the detection of drivers of the flood events. Eventually, we want to extend our analysis to long-term development of larger river systems.

Gunderson, L. H., and C. S. Holling, editors. 2002. *Panarchy: understanding transformations in human and natural systems.* Island Press, Washington, D.C., USA.