Geophysical Research Abstracts
Vol. 21, EGU2019-18467, 2019
EGU General Assembly 2019

DATALAKES - data platform and stochastic Bayesian forecasting for Swiss lakes using supercomputers

Artur Safin (1), Damien Bouffard (1), Fotis Georgatos (2), Eric Bouillet (2), Fernando Perez Cruz (2), Alfred Wüest (1,3), Siddhartha Mishra (4), and Jonas Sukys (1)

(1) Eawag, Siam, Switzerland (artur.safin@eawag.ch), (2) SDSC, Switzerland, (3) EPF Lausanne, Switzerland, (4) ETH Zurich, Switzerland

DATALAKES is a multi-disciplinary and multi-institutional project involving Swiss Data Science Center, Eawag, ETH Zurich, and EPF Lausanne.

As Chateau d'Eau of Europe, Switzerland requires a scientifically grounded lake management. Lakes are often misrepresented uniformly in blue, hiding rich spatio-temporal dynamics. Past researches have focused on vertical fluxes and structures. Today's additional challenge is to quantify horizontal transport and mixing processes influencing greenhouse gases emission (carbon and methane), oxygen depletion, harmful algae bloom, among many other ecologically relevant phenomena. New instrumentation, such as in situ observation platforms, remote sensing and computational resources enable the investigation of the temporal evolution of the environment's spatial heterogeneity. Yet, currently, environmental scientists still rarely manage to handle in a global integrated way these data for their studies. A key prerequisite are 3D numerical simulations capable of delivering forecast with uncertainties. However, required in situ observations, remote sensing and computational resources are under-utilized.

The first milestone of the DATALAKES project is to create a centralized data platform including in situ acquisition, storage, curation, patching, visualization, and extraction frameworks, together with an accessible online interface for visualization of historical data, future predictions, and allow user friendly online data extraction. Besides the end-to-end data platform, data mining techniques will be applied to analyze patterns and improve parameterization of deductive models, for instance, to provide a data-driven model of input processes (skin-to-bulk). The second milestone focuses on development of a scalable compute cluster enabled framework for predictive lake simulations that allow us to calibrate and predict lake dynamics. High performance computing will allow detailed quantification of uncertainties using Bayesian inference and modern Markov Chain Monte Carlo methods with multi-level variance reduction. The third milestone consists in demonstrating the benefit of data science and of the data platform for environmental lake study.

The DATALAKES project will provide a new data platform for the processing of massive amounts of hydrological and ecological data that will empower scientists, stakeholders, and citizens to view lakes as a dynamic system. The system will be first validated for Lake Geneva. This step is a prerequisite for future initiation of large interdisciplinary work. We envision a platform for monitoring lakes dynamics, including reliable weekly forecasts of three-dimensional lake ecological and physical states with hourly time series and associated uncertainties. Exploration of the dataset (e.g. machine learning, theory-based data science etc.) will allow more accurate water constituents budgets, including greenhouse gas emission, oxygen balance, harmful algae bloom, and fates of pollutants. We also hope to incite new collaborations between scientists, stakeholders, and citizens, and aim to further improve scientifically grounded management of water resources in Switzerland.