MSc. thesis project:
Model-based machine learning for air quality forecast

This is an MSc. thesis project for a student who is interested in applicating machine learning and numerical modeling to improve the accuracy of air quality forecast.

1 Introduction

Air pollution is one of the most important environmental issues of our time. For instance, according to a report by the World Health Organization (WHO, 2016), the passing away of one out of every nine persons is related to air pollution. Next to life and death, air pollution also causes great damage to economy. A dust storm or heavy smog with low visibility can cause a severe disruption of air traffic operations.

1.1 Chemical transport models

Over the last thirty years, large efforts have been spent in developing numerical atmospheric models in order to produce accurate air quality forecasts. Traditionally, the so-called chemical transport model (CTM) has been widely used to forecast the air quality index, such as Lotos-Euros (Manders et al., 2017) which is currently used for operational air index prediction in the Netherlands. These CTMs generally adopt (1) physical principles and (2) statistical methods to model the emission, advection, diffusion, and deposition. However, the accuracy of the CTMs is strongly affected by the model parametrization errors and the emission inventories. Here we note already that a timely update of the emission inventories is an essential prerequisite for an acceptable air quality forecast.

1.2 Data-driven machine learning

The advances in sensor technologies and the continuously decreasing costs of electronic devices have made large scale measurements feasible. A combination with the ever increasing power of computing platforms has led to a new paradigm in the computational and statistical methods for processing and analyzing data (Hey et al., 2009). It is collectively referred to as data science. Data-driven machine learning methods are nowadays able to deal with issues such as local refinement. However, current knowledge is not sufficient to formulate them into a (partial differential) equation. Therefore, data-driven machine learning techniques have been applied and they showed us some successes in improving relevant air quality predictions. Examples of using machine learning in atmospheric modeling have shown remarkable performances in a number of situations see (Li et al., 2016; Fan et al., 2017; Li et al., 2017;

Chen et al., 2018). Their results demonstrate that in some cases data-driven machine learning approaches are able to produce results with a high accuracy. However, we have to admit that the notion of a black-box application within data science has so far met only limited success (e.g., (Caldwell et al., 2014; Lazer et al., 2014)). Currently, we see in air quality forecasting research that the majority of the machine learning tools are data-driven and the knowledge about physical laws does not play any role of importance. As our starting point we put forward that scientific problems are often under-constrained in nature as the state space (the degree of freedom) is much larger than the training samples (observations). For example, the number of state variables in an atmospheric model is outnumbering the observations by far, because for a numerical model with millions or even billions grid points it is impossible to perform accurate measurements at every grid point and every time step.

1.3 Theory-based machine learning

Recently, several research groups have started to study the combination of physics and theory in data-driven machine learning models (Keller et al., 2017; Karpatne et al., 2017; Jia et al., 2018). An example is attempting to enforce physical consistency (e.g., conservation of mass and energy) through adding a regularization term in the loss function. It has resulted in more consistent output.

However, the parameterization of physical rules into the loss function in machine learning is also a complex work. For the atmospheric modeling the states of which are involved in many processes and have great spatiotemporal variability, it will be even more challenging and required the developer who are well familiar with the modeling techniques.

2 Model-based machine learning system

The research proposed here is to designed a model-based machine learning system for air quality forecast with a high feasibility. Not like the option 1 as shown in Fig.1 which combines the physics into the machine learning models by adding the regularizations. This research is to explore the possibility of combine the available chemical transport model with current data driven machine learning system. Actually the former one are considered as a good representation of the physics, and the output of which can be mapped into the observation space easily required by machine learning. As shown in Fig.1 option 2, the final system could use a CTM for generating output which is then used as input for a machine learning system.

The work includes the design of the new machine learning architecture, the sensitivity test of chemical transport model simulation into machine learning training model and also the optional configurations.

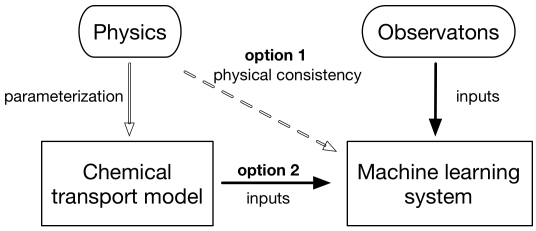

Figure 1. The combination of chemical transport model and machine learning system.

3 What to do for the model-based machine learning air quality forecast system

In this project, exploration on combining physical-based CTMs with data-driven machine learning will be carried out over the East Asian areas, mainly the China, over which a rich observation data are available, e.g., ground-based and satellite-based air quality measurements. A CTM (Lotos-Euros) is available to provide the daily air quality forecast over the test region. The study can comprise the following steps:

1: Learn about the air quality components involved processes and the Lotos-Euros model;

2: Use machine learning techniques, e.g., LSTM (long short term memory) neural network to predict the air quality index (only-data driven);

3: Combine the air quality forecasts from Lotos-Euros with the data-driven machine learning system, analyze the sensitivity of the new input features on the estimated forecasts, update the emission inventory that drives the CTM (optional);

4: Explore different configurations of the hybrid system, investigate the influence of the hyperparameters on the performance of the predictions.

References

Caldwell, P. M., Bretherton, C. S., Zelinka, M. D., Klein, S. A., Santer, B. D., and Sanderson, B. M.: Statistical significance of climate sensitivity predictors obtained by data mining, Geophysical Research Letters, 41, 1803–1808, https://doi.org/10.1002/2014GL059205, 2014.

Chen, G., Li, S., Knibbs, L. D., Hamm, N. A. S., Cao, W., Li, T., Guo, J., Ren, H., Abramson, M. J., and Guo, Y.: A machine learning method to estimate PM2.5 concentrations across China with remote sensing, meteorological and land use information, Science of The Total Environment, 636, 52–60, https://doi.org/10.1016/j.scitotenv.2018.04.251, 2018.

Fan, J., Li, Q., Hou, J., Feng, X., Karimian, H., and Lin, S.: A Spatiotemporal Prediction Framework for Air Pollution Based on Deep RNN, ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, IV-4/W2, 15–22, https://doi.org/10.5194/isprs-annals-IV-4-W2-15-2017%7D, 2017.

Hey, T., Tansley, S., and Tolle, K.: The Fourth Paradigm: Data-Intensive Scientific Discovery, Microsoft Research, 2009.

Jia, X., Karpatne, A., Willard, J., Steinbach, M., Read, J., Hanson, P. C., Dugan, H. A., and Kumar, V.: Physics Guided Recurrent Neural Networks For Modeling Dynamical Systems: Application to Monitoring Water Temperature And Quality In Lakes, 2018.

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., and Kumar, V.: Theory-guided Data Science: A New Paradigm for Scientific Discovery from Data, IEEE Transactions on Knowledge and Data Engineering, 29, 2318–2331, https://doi.org/10.1109/tkde.2017.2720168, 2017.

Keller, C. A., Evans, M. J., Kutz, J. N., and Pawson, S.: Machine learning and air quality modeling, 2017 IEEE International Conference on Big Data (Big Data), pp. 4570–4576, https://doi.org/10.1109/BigData.2017.8258500, 2017.

Lazer, D., Kennedy, R., King, G., and Vespignani, A.: The Parable of Google Flu: Traps in Big Data Analysis, Science, 343, 1203–1205, https://doi.org/10.1126/science.1248506, 2014.

Li, X., Peng, L., Hu, Y., Shao, J., and Chi, T.: Deep learning architecture for air quality predictions, Environmental Science and Pollution Research, 23, 22 408–22 417, https://doi.org/10.1007/s11356-016-7812-9, 2016.

Li, X., Peng, L., Yao, X., Cui, S., Hu, Y., You, C., and Chi, T.: Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation - ScienceDirect, Environmental Pollution, 231, 997–1004, https://doi.org/https://doi.org/10.1016/j.envpol.2017.08.114, 2017.

Manders, A. M. M., Builtjes, P. J. H., Curier, L., Denier van der Gon, H. A. C., Hendriks, C., Jonkers, S., Kranenburg, R., Kuenen, J., Segers, A. J., Timmermans, R. M. A., Visschedijk, A., Wichink Kruit, R. J., Van Pul, W. A. J., Sauter, F. J., van der Swaluw, E., Swart, D. P. J., Douros, J., Eskes, H., van Meijgaard, E., van Ulft, B., van Velthoven, P., Banzhaf, S., Mues, A., Stern, R., Fu, G., Lu, S., Heemink, A., van Velzen, N., and Schaap, M.: Curriculum vitae of the LOTOS-EUROS (v2.0) chemistry transport model, Geoscientific Model Development, 10, 4145–4173, https://doi.org/10.5194/gmd-10-4145-2017, http://www.geosci-model-dev-discuss.net/gmd-2017-88/, 2017.

WHO: Ambient air pollution: a global assessment of exposure and burden of disease, 2016.