Data Science in Climate and Climate Impact Research

Conceptual Issues, Challenges, and Opportunities

Virtual Workshop at ETH Zurich
August 20 and 21, 2020

Organizers:
Dr. Christoph Baumberger
Prof. Dr. David N. Bresch
Dr. Benedikt Knüsel
Prof. Dr. Reto Knutti
Marius Zumwald

ETH zürich

75 NRP Big Data
National Research Programme

1. Contents

2. Time Table

Zoom links

The keynotes, the social activity, and the talks of Parallel Session 1 will take place in the following zoom meeting room:
https://ethz.zoom.us/s/97080726192

The talks of Parallel Session 2 will take place in the following zoom meeting room:
https://ethz.zoom.us/s/97186438549

Thursday, August 20, 2020

(all times are indicated according to Central European Time)

Time	Parallel Session 1	Parallel Session 2
13:00 - 13.15	**Welcome** Reto Knutti	
13:15 - 14:15	Title tbc Markus Reichstein	
	Remote Sensing	**Machine Learning and Transparency**
14:30 - 15:00	*Advances and limitations in the use of satellite imagery for deforestation and degradation monitoring and reduction in tropical forests* Federico Cammelli, Owen Cortner, Janina Grabs, Samuel Levy, Radost Stanimirova, Rachael Garrett	*Transparency, Interpretability and Data Availability: Key Challenges in Tackling Climate Change with AI* Joyjit Chatterjee, Nina Dethlefs
15:00 - 15:30	*Towards Data-Informed Climate Sciences - Leveraging Machine Learning Inferences of Satellite Observations* Srija Chakraborty	*Exploring deep neural networks for probabilistic postprocessing of NWP wind forecasts in complex terrain* Daniele Nerini, Max Hürlimann, Lionel Moret, Jonas Bhend, Mark Liniger
15:30 - 16:00	*Planetary Scale Location Insights serving Climate Adaptation* Gopal Erinjippurath	*The Importance of Neural Network Interpretation Techniques for Climate and Weather Science* Amy McGovern, Ryan Lagerquist, Elizabeth Barnes, Imme Ebert-Uphoff
16:00 - 16:30	Break	
16:30 - 17:30	**Evaluating data: a fitness-for-purpose view** Wendy Parker	
from 18:00	Social Activity	

Friday, August 21, 2020

(all times are indicated according to Central European Time)

Time	Parallel Session 1	Parallel Session 2
	Causality & Scientific Understanding	**Climate Social Sciences**
08:30 - 09:00	*Response-Guided Learning to boost S2S Forecasting* Sem Vijverberg, Dim Coumou	*Opportunities of machine learning in agricultural insurance* Tobias Dalhaus, Thomas Heckelei, Robert Finger
09:00 - 09:30	*The Impact of Statistics and Machine Learning on Understanding in Climate Modeling* Julie Jebeile, Vincent Lam, Tim Raez	*A User Study of Perceived Carbon Footprint* Victor Kristof, Lucas Maystre, Matthias Grossglauser, Patrick Thiran
09: 30 - 10:00	*Causal Networks as a framework for climate science to improve process understanding* Marlene Kretschmer, Ted Shepherd	*Projecting Downscaled Social and Behavioral Impacts from Climate Change Using Mobile Devices* Kelton Minor, Andreas Bjerre-Nielsen, Jonas Skjold Raaschou-Pedersen, Sune Lehmann, David Dreyer Lassen
10:00 - 10:30	Break	
10:30 - 11:30	**Synthesis of the Project: "Combining Theory with Big Data"** within the National Research Programme "Big Data" Benedikt Knüsel, Marius Zumwald	
	Uncertainty in Observational Data and Modeled Outputs	**Domain-Specific Background Knowledge and Machine Learning**
11:45 - 12:15	*Sparse principal component analysis as a tool to explore heterogeneous datasets from multidisciplinary field experiments* Sebastian Landwehr, Michele Volpi, Fernando Perez-Cruz, Julia Schmale	*Stochastic generation of climate and weather data fields with generative adversarial networks* Jussi Leinonen, Alexis Berne
12:15 - 12:45	*Addressing uncertainty in climate models data and an overview of application of data science in climate studies* Titas Ganguly, Dhyan Singh Arya	*How to combine domain knowledge with the capacity of machine learning for discovery?* Eniko Szekely
12:45 - 13:15	*Towards a generalized framework for missing value imputation of fragmented Earth observation data* Verena Bessenbacher, Lukas Gudmundsson, Sonia Seneviratne	*The wrong dichotomy: Data science as supplementing, rather than displacing other methods* Dominik Fitze
13:15 - 14:30	Break	
	Interdisciplinarity and Scientific Practice	**Modeling and Representation**
14:30 - 15:00	*Challenges and Opportunities in the Publication of Global-scale, High-resolution Climate Model Outputs* Ionut Iosifescu Enescu, Gian-Kasper Plattner, Dirk Nikolaus Karger, David Hanimann, Dominik Haas-Artho, Martin Hägeli, Niklaus E. Zimmermann, Konrad Steffen	*(How) can machine learning be used for predictive modeling of the Earth system?* Benjamin Stocker

15:00 - 15:30	*Situated knowledge and climate services: miscellaneous scales and levels of interpretation, using physical observations and data science* Melissande Machefer, Arthur Brun	*Building a Seasonal Earth System Model Emulator for local temperature* Shruti Nath, Quentin Lejeune, Lea Beusch, Carl Friedrich Schleussner, Sonia Seneviratne
15:30 - 16:00	*Multilingual Structured Climate Research Data in Wikidata - The Community Perspective* Cristina Sarasua, Daniel Mietchen	*Can Machines Learn Convection? The epistemic implications of machine-learning parameterizations in climate science* Suzanne Kawamleh
16:00 - 16:30	*Multilingual Structured Climate Research Data in Wikidata - The Data Perspective* Daniel Mietchen, Cristina Sarasua	*The role of machine learning in site-specific wind turbine power curve prediction* Sarah Barber
16:30 - 17:00	Break	
17:00 - 18:00	**Prospects for data science to advance the study of social climate impacts** Nick Obradovich	
18:00 - 18:30	**Closing Words** Reto Knutti & David Bresch	

3. Keynotes and Plenaries

P1

Machine-learning-model-data-integration for a better understanding of the Earth System

Markus Reichstein

The Earth is a complex dynamic networked system. Machine learning, i.e. derivation of computational models from data, has already made important contributions to predict and understand components of the Earth system, specifically in climate, remote sensing and environmental sciences. For instance, classifications of land cover types, prediction of land-atmosphere and ocean-atmosphere exchange, or detection of extreme events have greatly benefited from these approaches. Such data-driven information has already changed how Earth system models are evaluated and further developed. However, many studies have not yet sufficiently addressed and exploited dynamic aspects of systems, such as memory effects for prediction and effects of spatial context, e.g. for classification and change detection. In particular new developments in deep learning offer great potential to overcome these limitations.

Yet, a key challenge and opportunity is to integrate (physical-biological) system modeling approaches with machine learning into hybrid modeling approaches, which combines physical consistency and machine learning versatility. A couple of examples are given with focus on the terrestrial biosphere, where the combination of system-based and machine-learning-based modelling helps our understanding of aspects of the Earth system.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., Prabhat, 2019. Deep learning and process understanding for data-driven Earth system science. Nature 566, 195-204.

P2

Evaluating data: a fitness-for-purpose view

Wendy Parker

There is a tendency to think that data and data models are ‘good’ to the extent that they mirror the world. We suggest an alternative view that calls for evaluating data and data models according to their adequacy or fitness for particular purposes. We discuss some implications of adopting this fitness-for-purpose view and show how such a view aligns with salient features of practice, especially the iterative reuse and repurposing of data.

P3

Synthesis of the Project: “Combining Theory with Big Data”

Benedikt Knüsel, Marius Zumwald

In recent years, the ability to gather and store information has increased dramatically, and the ability to make use of these increasing volumes of data has improved. This advent of big data has opened up new opportunities for scientific research, including for research on climate change. The NRP75 project «Combining Theory with Big Data» explored new opportunities and challenges associated with big data in climate research. It did so by combining a philosophy of science perspective with applied work aiming to predict urban temperature distribution in high-resolution. In this synthesis talk, we outline some of the central findings from this project. We specifically discuss how the uncertainty of datasets and data-driven models may be assessed, and how data-driven modeling may be employed by researchers who aim not only at predicting but also at understanding. Based on the philosophical and applied work, we also discuss how and why domain-specific background knowledge is crucial for applying new forms of data and data-driven models in scientific contexts.

P4

Prospects for data science to advance the study of social climate impacts

Nick Obradovich

What have we learned from the use of data science in the study of the social impacts of climate? What are some of the most important outstanding questions? Why do these questions even matter? And how might data science tools and techniques enable us to answer them? In this talk I will outline some priorities, potential pitfalls, and promising prospects for the use of data science in estimating the future social impacts of climate change.

4. Abstracts of Talks

Remote Sensing

Thursday, Parallel Session 1, 14:30 – 16:00

1

Advances and limitations in the use of satellite imagery for deforestation and degradation monitoring and reduction in tropical forests.

Federico Cammelli[1], Owen Cortner[1], Janina Grabs[1], Samuel Levy[1], Radost Stanimirova[2], Rachael Garrett[1]

[1]ETH, [2]Boston University

The continued deforestation and degradation of primary forests and biodiversity hotspots in tropical areas are crucial challenges in the fight against global warming and biodiversity loss. The majority of forest loss is due to the expansion of commodity agriculture and commercial forestry operations, including soy, beef cattle, palm oil, timber, and pulp and paper production. Satellite imagery has become an important tool for practitioners to monitor forest cover loss in real time, assess the compliance of actors with policies that prohibit deforestation and degradation, and select priority non-compliant actors or regions to engage with. Remotely sensed observations are also the primary data source to estimate changes in the rates of forest change, its causes, and the effectiveness of forest policies. The specifics of how remotely sensed observations are utilized in science and practice determines the ecological and social outcomes of forest governance interventions. This paper provides a review of the existing literature and governmental and non-governmental programs that utilize remote sensing for land change detection and monitoring to: 1) Assess advances in the use of satellite-based observations and data science tools in tropical forest research and practice and 2) Identify common pitfalls associated with satellite data use that undermine the rigor and generalizability of analyses based on remotely sensed data. These pitfalls involve low primary data quality (especially regarding imagery dates and time series) and a lack of best practice consensus for reporting and comparing outcomes, bias, and uncertainty as a result of differences in definitions of deforestation and degradation, sampling approaches, and temporal and spatial reporting and attribution decisions. We conclude by providing recommendations to increase the effectiveness of using satellite-based monitoring systems for coupled human-natural system analysis and forest governance.

2

Towards Data-Informed Climate Sciences - Leveraging Machine Learning Inferences of Satellite Observations

Srija Chakraborty[1]

[1]NASA Goddard Space Flight Center

Data acquired by the ever-increasing Earth observing satellites have created a rich repository comprised of fine spatial, spectral and temporal resolution observations collected by sensors with varying modalities. Uniform data acquisition on a global scale, makes such datasets suitable for stuyding different Earth system components, retrieving its geophysical parameters, monitoring short- and long-term trends, and analyzing these variations to understand changes in climate, its drivers and impact. However, the large data volume and the complexity of the patterns, necessitates the utilization of machine learning (ML) algorithms to extract meaningful information from these observations.

Additionally, for gaining accurate insights from ML models of satellite observations, four aspects are identified that should be incorporated in the analysis pipeline to increase the impact of ML in climate sciences and policies.

Firstly, incorporating domain knowledge and physical models while training ML models is essential to enforce constrains and relationships between parameters that may not be inferrable solely from the data. Knowledge-guided models are more reliable as the predictions satisfy domain-specific equations. Moreover, it is crucial to explore interpretable machine learning for increasing a scientists' confidence on the models and for deciphering the most relevant features for the task.

Secondly, ML models are dependent on data quality. Varying acquisition conditions lower the quality of retrieved geophysical parameters which are used for model training, thereby deteriorating the model quality and predictions. Careful consideration of data quality and simulations of uncertainty are necessary as the model outputs are utilized for further analyses.

Thirdly, evaluation metrics should be tailored to analyze performance on the task at hand at diverse geographic locations by assessing the results with statistical measures and on its acceptability to domain-experts.

Finally, all abovementioned factors will require interdisciplinary collaboration and domain-expert feedback at every stage of the ML pipeline for data-informed climate science and policy.

3

Planetary Scale Location Insights serving Climate Adaptation

Gopal Erinjippurath[1]

[1] *Sust Global*

The recent explosion of new data streams from varied satellite streams necessitates the use of automation techniques that allow for processing and refining raw sensor data to streams of foundational feature information for business insights and geosciences research. Deep learning provides new avenues for deriving insights from imagery. Recent approaches allow for feature detection and localization over varied context and different imaging conditions. Overhead satellite imagery is varied and rich, but fundamentally different from other data sources and introduces unique challenges towards scalable location insights.

We will explore the use of modern deep learning approaches to object detection, semantic segmentation towards feature extraction and meaningful change detection in satellite imagery. We identify techniques towards sampling data, leveraging heterogeneous datasets, performance benchmarking and tuning models towards generalized performance. We will walk through the technical challenges involved with serving results from these deep learning models in a scalable manner along with essential metadata for interpretability and usability for our customers. We then explore approaches with human in the loop towards improved performance over time towards scalable location-based intelligence served across the entirety of the Earth’s landmass and discuss the challenges and opportunities at the frontiers of geospatial data and location-based insights. Specific applications of using such derived data streams along with climate risk indicators to serve insights towards climate adaptation will be presented.

By imaging the Earth every day at 3.7 m resolution and enabling on-demand follow up imagery at 72 cm resolution, Planet offers a uniquely valuable dataset for creating datasets for imagery analytics over varied context. We use the above approaches towards creating foundational data feeds serving new applications – and how this new data feed can be

spatially joined with other data sources to serve location insights, that help us explore and better understand climate adaptation.

Machine Learning and Transparency

Thursday, Parallel Session 2, 14:30 – 16:00

4

Transparency, Interpretability and Data Availability: Key Challenges in Tackling Climate Change with AI

Joyjit Chatterjee[1], Nina Dethlefs[1]

[1]*University of Hull*

With growing natural disasters, rise in carbon emissions and faltering ecosystems, the need for furthering research in climate change has become integral. Recent studies have shown that data science can play a vital role in better understanding natural phenomena and discovering novel insights. Although no silver bullet, machine learning (ML) has been successfully utilised in an array of applications, ranging from prediction and assessment of droughts and floods, energy control in grids, water quality modelling, operations & maintenance (O&M) of renewable energy sources such as wind and solar energy etc. However, the existing studies suffer from 2 prime challenges:

(1) Lack of data availability - domain specific information e.g. from wind turbines, is often commercially sensitive, making it difficult to procure large amounts of useable data - especially new kinds of data which can possibly generate significant new insights. Transfer learning techniques can help learn from little or no labelled data, ensuring accuracy and helping algorithms to generalise better.

(2) The black-box nature of (deep) ML models makes them suffer from the problem of transparency, wherein, although predictions can often be made with high accuracy, confidence and trust in the model decisions is difficult. A human intelligible diagnosis of when, why, what and how a model performs (or not) is essential. Hybrid ML techniques can bridge the gap between transparency and accuracy, and causal inference can help discover hidden insights from data. Natural language generation can further help in generating informative reports and descriptions of natural disasters and O&M strategies for renewable energy sources.

We believe that there is enormous opportunity for the data science community to pursue research to tackle some of these challenges in ensuring reliable decision making and envisage that making data-driven decision support systems intelligent and transparent would have a significant impact in tackling climate change.

5

Exploring deep neural networks for probabilistic postprocessing of NWP wind forecasts in complex terrain

Daniele Nerini[1], Max Hürlimann[1], Lionel Moret[1], Jonas Bhend[1], Mark Liniger[1]

[1]*Federal Office of Meteorology and Climatology MeteoSwiss*

Despite the many success stories related to the physical approach of numerical weather predictions (NWPs), the accurate forecasting of surface winds and their corresponding uncertainty in complex terrain remains an important challenge. Even for kilometer-scale NWP, many local topographical features remain unaccounted, often resulting in biased forecasts with respect to local weather conditions.

Through statistical postprocessing of NWP, such systematic biases can be adjusted a posteriori using wind measurements. However, for unobserved locations, these approaches fail to give satisfying results. Indeed, the complex and nonlinear relationship between model error and topography calls for more advanced techniques such neural networks (NN).

Furthermore, the prevalence of aleatoric uncertainties in wind forecasts demands the adoption of a probabilistic approach where the statistical model is not only trained to predict an error expectation (the bias), but also its scale (standard deviation). In this context, the model must be trained and evaluated using a proper scoring rule (Gneiting and Raftery 2007).

In an interdisciplinary effort between meteorology and computer science, we developed a machine learning application to efficiently handle very large datasets (order of TBs), train various probabilistic NN architectures, and test multiple combinations of predictors. We used a game theoretic approach (Lundberg and Lee 2017) to explain the predictions from the deep neural networks and thus maintain a certain level of interpretability within data-driven models.

As a result, we were able to improve the quality of the NWP model output not only at the location of reference measurements, but also at any given point in space, and for forecasts up to 5 days into the future. More importantly, the results underline that the combination of physical models with a data-driven approach opens new opportunities to improve weather forecasts and in particular weather warnings.

6

The Importance of Neural Network Interpretation Techniques for Climate and Weather Science

Amy McGovern[1], Ryan Lagerquist[1], Elizabeth Barnes[2], Imme Ebert-Uphoff[2]

[1]University of Oklahoma, [2]Colorado State University

In this talk we highlight important uses of neural network visualization techniques for climate and weather applications. As neural networks/deep learning become more widely used in these domains, scientists want to gain insights into the strategies used by the neural network and understand the physical relationships that a model has learned about the tasks being studied. Such understanding has several benefits: 1) It increases trust in the neural network model and 2) We can discover new science by having the neural network discover not-yet-known relationships.

We present multiple approaches for model interpretation of neural networks that we have found particularly useful and illustrate them for three different applications, namely for predicting hailstorms, predicting tornadoes, and to identify spatial patterns of climate change based on global, annually averaged temperature maps. Network-based interpretation and visualization techniques are varied in how they highlight what the model has learned. Some approaches, including saliency maps, Grad-CAM, and Layer-wise Relevance Propagation (LRP), highlight specific areas of the inputs that the network focuses on to make its prediction, and can also be used to visualize the intermediate layers of the network. We demonstrate the use of saliency maps and Grad-CAM for the hailstorm and tornado networks, and the use of LRP to identify spatial patterns of climate change. We also present another approach, backwards optimization, which uses gradient descent to create a synthetic input that activates the neural network in a certain way. We demonstrate the use of backward optimization in the context of tornado prediction to create a thunderstorm that

maximizes predicted tornado probability, and in the context of finding patterns of climate change to identify ubiquitous climate model biases. Finally, we discuss lessons learned about the overall process of identifying physically meaningful results from neural networks for climate and weather applications.

Causality and Understanding

Friday, Parallel Session 1, 08:30 – 10:00

7

Response-Guided Learning to boost S2S Forecasting

Sem Vijverberg[1], Dim Coumou[1]

[1]*Institute for Environmental Studies, VU Amsterdam*

Seasonal to sub-seasonal (S2S) predictability could provide societies with valuable information on weather-related risk, allowing decision-makers to initiate early warning action plans and to optimize resource management. Teleconnections can be an important source of predictive skill, yet dynamical models often fail to represent teleconnections accurately. The model development community can therefore benefit from a better understanding of the underlying physical drivers. One can also directly use the statistical information deduced from physical drivers and merge it with information from dynamical models, i.e. creating a hybrid forecast model. The dominant physical drivers differ for different geographical regions and different timescales. Machine learning might help to extract the most important physical drivers from multi-dimensional climate data by reducing the dimensionality while maintaining all relevant information of all potential drivers.

We are developing a rigorous data-driven framework that enables automatic detection of interpretable physical drivers on S2S timescales. The framework is suited to cast this information into a statistical model while attempting to minimize overfitting issues, even though we are limited by the number of independent data-points on these timescales. From econometrics we known that forecast skill does not inform about the causal structure. Therefore, the framework uses statistical learning in conjunction with an advanced causal inference technique.

So far, this framework has been successfully applied to study and predict the Polar vortex variability and U.S. heatwaves. In future work, we hope to improve the dimensionality reduction step and introduce wavelet-like transformations on the features to extract the signal from the noise.

8

The Impact of Statistics and Machine Learning on Understanding in Climate Modeling

Julie Jebeile[1], **Vincent Lam**[1], Tim Raez[1]

[1]*University of Bern*

In this paper, we investigate how the use of statistical methods and machine learning techniques affects our ability to understand in climate modeling. Prima facie, the main goal of climate modeling is to provide projections in view of decision-making with respect to climate change, while understanding is secondary. As a consequence, it could be thought that the use of machine learning techniques in climate modeling is unproblematic, because these techniques do considerably enhance our predictive abilities, despite the fact that machine learning models are black boxes.

We argue for a more nuanced position. Based on a multidimensional and graded notion of understanding, we maintain that understanding is indispensable to appropriately evaluate

climate models. To support our position, we first articulate four criteria for understanding. Understanding a climate model in a particular context does not only involve empirical adequacy, but also our grasp of processes producing outputs, physical consistency of outputs, and the scope of the validity of the climate model. What is more, these criteria are not categorical, but come in degrees. We then put these four criteria to work in two cases of climate modeling. In the first case, we investigate how the use of statistical downscaling in regional climate modeling affects understanding; in the second case, this is contrasted with the use of deep neural networks as an alternative to super-parametrization in a global circulation model.

The main upshot of the paper is a twofold continuity of understanding. First, the use of machine learning will decrease understanding along some dimensions; however, the same tendencies can also be observed for more traditional statistical methods. Second, there is a tradeoff between an increase in empirical adequacy (with respect to the validated physical domain) and a decrease along the other three dimensions.

9

Causal Networks as a framework for climate science to improve process understanding

Marlene Kretschmer[1], Ted Shepherd[1]

[1]*University of Reading*

In the light of ongoing anthropogenic climate change and associated risks, supporting regional decision making should be a guiding principle of climate research. However, seasonal forecast models only have low skill and climate models often give inconclusive results about regional aspects of climate change. One major source of uncertainty are dynamical drivers in the climate system, such as storm tracks or the stratospheric polar vortex, which are not well understood theoretically and where models show diverse responses.

The recent hype of machine learning promises data-driven solutions to these issues. While data-centric methods such as deep learning have and certainly will make notable contributions to the earth sciences, their power lies in their ability to efficiently describe complex relationships present in the data. There is reason to doubt whether these methods can, on their own, deal with the sort of epistemic uncertainty described above. Moreover, machine learners and climate scientists often lack a common language, making successful collaboration still difficult. In particular, climate scientists are trained to think in terms of causal relationships, whereas machine learning is mostly descriptive (i.e. correlational) and does not explicitly incorporate domain knowledge.

Here we call for the use of causal networks in climate science as a framework to overcome some of these challenges. We argue that causal networks are a simple yet powerful tool to translate qualitative expert knowledge about physical processes into mathematical objects, to gain quantitative information about the role of these processes through applying the rules of causal inference.

Climate Social Sciences

Friday, Parallel Session 2, 08:30 – 10:00

10

Opportunities of machine learning in agricultural insurance

Tobias Dalhaus[1], Thomas Heckelei[2], Robert Finger[1]

[1]*ETH Zürich,* [2]*University of Bonn*

Weather extremes affect agricultural production and thus threaten global food security. Traditionally, insurances provide payouts to farmers to reduce the financial exposure to these risks. Classical indemnity based insurances, where loss adjusters visit the affected farms to quantify the losses, cannot be applied at large scale for increasingly important perils such as drought and heat, due to information asymmetries between farmers and insurers. Weather index insurances complement insurance solutions by providing data driven payouts based on measurable weather conditions (i.e. a weather index; such as the rainfall at a weather station) without requiring on-farm yield measurements. This payout is determined by a function that is usually parametrized using regression techniques that are informed by observed farm-level yield records and weather index data. For index building, data sources like satellites, weather stations, phenology reporters or farm surveys are pulled together. The quantity and quality of available data for these purposes is increasing massively. Based on estimated relationships based on past observations the payout function is estimated and future insurance payouts are based weather index values only.

Machine learning approaches are currently not used in this process, but we here argue that future research on weather index insurance design should incorporate machine learning approaches that are expected to better predict yields based on environmental conditions. Machine learning is designed to get a prediction of an outcome variable (here crop yields) based on one or several input variables (here weather and other environmental conditions). Especially, deep neural network techniques can help to incorporate the vast amounts of data on various environmental conditions and their often non-linear interactions to predict crop yields. Thus machine learning techniques can deliver spatially explicit yield predictions. This potentially helps to better understand agricultural weather risks in space and time, which in turn can inform index insurance payouts.

11

A User Study of Perceived Carbon Footprint

Victor Kristof[1], Lucas Maystre[1], Matthias Grossglauser[1], Patrick Thiran[1]

[1]*Ecole Polytechnique Fédérale de Lausanne*

To put the focus on actions that have high potential for emission reduction, we must first understand whether people have an accurate perception of the carbon footprint of these actions. If they do not, their efforts may be wasted. We aim at modeling how people perceive the carbon footprint of their actions, which could guide better-informed climate communication.

Consumers and citizens repeatedly face multiple options with varying environmental impact. Except for a handful of experts, nobody is able to estimate the absolute quantity of CO2 emitted by their actions, say flying from Zurich to Berlin. Most people, however, are aware that taking the train for the same trip would release less CO2. Hence, in the spirit of social-

psychology studies, we posit that the perception of a population can be probed by simple pairwise comparisons: Instead of asking difficult questions about each action and averaging the answers, we ask simple questions in the form of comparisons and design a non-trivial statistical model to estimate the perception.

Our contributions are as follows. First, we cast the problem of inferring a population's global perception from pairwise comparisons as a Bayesian linear regression. The Bayesian formulation of the model enables us to take an active-learning approach to maximize the information gained from each comparison, i.e., this enables us to select the actions to compare in a statistically significant way. Finally, we develop a Web platform to collect real data from users in Switzerland.

We perform an in-depth life cycle analysis of the carbon footprint of Swiss citizens to obtain a list of actions with ground-truth carbon footprint. We compare these true values to the perceived carbon footprint estimated by our model to reveal discrepancies (underestimation and overestimation) for some actions, suggesting that some individual mitigation efforts could be adjusted.

12

Projecting Downscaled Social and Behavioral Impacts from Climate Change Using Mobile Devices

Kelton Minor[1], Andreas Bjerre-Nielsen[1], Jonas Skjold Raaschou-Pedersen[1], Sune Lehmann[2], David Dreyer Lassen[1]

[1] *University of Copenhagen,* [2] *Technical University of Denmark*

The first and second order effects of global climate change are already being detected, attributed and addressed locally[1–5]. In recent years, the practice of downscaling climate models to bridge the void between projected global effects and regional impacts has provided an essential guiding reference for policy makers and adaptation planners[6,7]. However, beyond understanding probable physical impacts under different emissions scenarios, climate impact researchers and decision makers increasingly seek ever more localized estimates of expected climate impacts on the behavior, productivity and well-being of specific human social systems, institutions and communities[8]. Drawing on novel empirical tools from climate econometrics[9] and reality mining methods from computational social science[10,11], we link over two years of minute-to-minute social and behavioral data from mobile phones for a specific social system - the freshman cohort of the Technical University of Denmark - with high resolution meteorological and climatological data. Using the Copenhagen Networks Study[12] as an illustrative case, we discuss the prospects and perils of linking empirically-derived historical estimates of localized human environment relationships with downscaled and bias-corrected climate model output. Specifically, we highlight issues of measurement error across spatially and temporally aggregated data sets and discuss possible threats when extrapolating historical inferences to future time horizons and cohorts. We contend that careful consideration should be paid to both uncertainty in downscaled climate model output and in plausible social ecological developments when assessing the dependability of quantitative behavioral impact projections.

Uncertainty in Observational Data and Model Outputs

Friday, Parallel Session 1, 11:45 – 13:15

13

Sparse principal component analysis as a tool to explore heterogeneous datasets from multidisciplinary field experiments

Sebastian Landwehr[1], Michele Volpi[2], Fernando Perez-Cruz[2], Julia Schmale[1]

[1] *Extreme Environments Research Laboratory, École Polytechnique Fédérale de Lausanne, School of Architecture, Civil and Environmental Engineering, Lausanne, Switzerland*

[2] *Swiss Data Science Center, ETH Zurich and EPFL, Switzerland*

During the research cruise Antarctic Circumnavigation Expedition (ACE) during the Austral summer 2016/2017 a multidisciplinary team of researchers made observations of many oceanic and atmospheric variables in a wide range of environmental conditions and in a region of the world that is today still heavily undersampled. Overviewing and connecting the bounty of observations, which were sampled at different temporal resolutions ranging from seconds to several hours is both challenging and necessary to facilitate interdisciplinary research on the processes that connect life in the ocean with the concentration of trace gases and aerosols in the atmosphere.

Within the ACE-DATA - (Delivering Added-value To Antarctica) research project funded by the Swiss Data Science Center (SDSC) we have developed an approach based on sparse Principal Component Analysis (sparse PCA) that decomposes time series of observed variables (onboard measurements) into a set of latent variables (principal components) of much lower dimension. This allows to highlight common patterns in the along-track spatio-temporal variation of the observed variables. The joint variability of observations made by the different research groups, which become visible as activations of the latent variables, have sparked discussions and helped to invigorate the collaboration between the research projects. We further quantify the sensitivity of the results to spurious observations by means of a bootstrapping strategy. We will present our experiences and discuss the limits and potential of this method to support interdisciplinary research projects.

14

Addressing uncertainty in climate models data and an overview of application of data science in climate studies

Titas Ganguly[1], Dhyan Singh Arya[1]

[1]*Indian Institute of Technology Roorkee*

Uncertainties in climate model data stem from structural as well natural causes. According to literature the reduction of uncertainty is directly proportional to the number of models used in mathematical average ensemble. However, we show here that weighted average of five models has higher efficiency (in comparison to mathematical average) in reduction of uncertainty. Weighted average ensemble data was generated by a novel statistical methodology (based on Bayesian and Orthonormal distribution) using CMIP-5 models for eight Koppen climate zones of India. The uncertainty was calculated as the interquartile range of 2.5 and 97.5 percentile of a nonparametric PDF of projected anomalies of temperature and precipitation. It was seen that the ensemble data had minimum uncertainty for all grids for precipitation and for 72.67% grids for maximum temperature (RCP 4.5 and

8.5 scenarios). Similar results have been reported from data science backgrounds using climate models.

Data science techniques are developing with the exponential growth in data (4.4 zettabytes in 2013) and it naturally finds applications in data rich disciplines like climate science. Since it is not constrained by theory, data science can potentially address challenges like non-linearity and non-stationarity, parameterisation problems, multi-dimensionality etc., in climate studies. Application of data science in climate also has various constraints. The objectives and language of climate studies are rooted in natural sciences thus necessitating integration (of theory and data mining), both at design and interpretation levels, and not complete absence of theory. Data mining techniques are mostly based on the emergent attributes without considering the inherent characteristics of data itself (eg: spatio-temporal specificity and variability). Also assumptions of independence and homogeneity in data science techniques are rarely valid for climate data while sampling of extreme events pose challenges with existing data algorithms. Thus opportunities and challenges are abundant in application of data-science to climate studies.

15

Towards a generalized framework for missing value imputation of fragmented Earth observation data

Verena Bessenbacher[1], Lukas Gudmundsson[1], Sonia Seneviratne[1]

[1]*ETH Zürich*

The past decades have seen massive advances in generating Earth System observations. A plethora of instruments is, at any point in time, taking remote measurements of the Earth's surface aboard satellites. This has become invaluable to the climate science community. However, the same variable is often observed by several platforms with contrasting results and satellite observations have non-trivial patterns of missing values. Consequently, mostly only one remote sensing product is used simultaneously. This has led to a fragmentation of the observational record that limits the widespread use of remotely sensed land observations. We aim towards a generalized framework for mutually gap-filling global, high-resolution remote sensing measurements relevant for the terrestrial water cycle, focusing on soil moisture, land surface temperature and precipitation. To this end, we explore statistical imputation methods and benchmark them using a "perfect dataset approach", in which we apply the missingness pattern of the remote sensing datasets onto their matching variables in the ERA5 reanalysis data. Original and imputed values are subsequently compared. Our approach iteratively produces estimates for the missing values and fits a model in an expectation-maximisation alike fashion. This procedure is repeated until the estimates for the missing data points converge. The method harnesses the highly-structured nature of gridded covarying observation datasets within the flexible function learning toolbox of data-driven approaches. The imputation utilises (1) the temporal autocorrelation and spatial neighborhood within one dataset and (2) the different missingness patterns across datasets, i.e. the fact that if one variable at a given point in space and time is missing, another covarying variable might be observed and their local covariance could be learned. A method based on ridge regression has shown to perform best. This model will be applied to gapfill satellite data and create an inherently consistent dataset based exclusively on observations.

Domain-Specific Background Knowledge and Machine Learning

Friday, Parallel Session 1, 11:45 – 13:15

16

Stochastic generation of climate and weather data fields with generative adversarial networks

Jussi Leinonen[1], Alexis Berne[1]

[1]*Ecole Polytechnique Fédérale de Lausanne*

Data problems in atmospheric science commonly deal with spatial fields that have complex structure. Moreover, observations tend to be incomplete due to sparse spatial coverage, inadequate resolution, or uncertainties in the measurement process. This makes climate and weather attractive applications for deep learning, which is well suited to processing spatial fields with complex patterns. However, trying to predict such fields using typical neural networks tends to lead to regression to the mean, yielding blurred results that do not have the correct spatial structure. Moreover, predictive models do not generally describe the uncertainty of their predictions, while uncertainty quantification is critical in many climate applications.

Generative adversarial networks (GANs) can address the above-mentioned limitations by generating spatially realistic output fields stochastically, producing a distribution of solutions rather than a single answer, ideally (but in practice not always) converging to the real variability of the underlying data distribution. A straightforward GAN variant called the conditional GAN can be trained to generate solutions corresponding to a condition given as an input. These can be used for common problems in weather and climate data processing, such as generating physical fields from the corresponding in-situ and remote sensing observations, increasing the resolution of observed data, or predicting the time evolution of data fields. Other GAN variants can be used for unsupervised classification, enabling information extraction without manual labeling of training examples.

In this presentation, I will give an overview of GAN theory and the architecture of the neural networks needed to implement them. I will show case examples of my work with GANs so far, and discuss more generally which problems in climate science could (or already do) benefit from them. Furthermore, I'll discuss the current challenges for training GANs for weather and climate applications, and in validating and interpreting their results.

17

How to combine domain knowledge with the capacity of machine learning for discovery?

Eniko Szekely[1]

[1]*Swiss Data Science Center, ETH Zurich and EPFL*

Both machine learning and climate science require underlying knowledge: one on the algorithmic side and the other on the domain side. The aim of this talk is to discuss similarities and differences between the approaches taken in climate science and machine learning, and to question whether by combining them we could further advance our understanding of the climate. Besides prediction, one of machine learning's most attractive

features is its ability to discover unknown patterns and interactions in the data and to study them in a unified framework. Such interactions often occur between phenomena that evolve at different temporal and spatial scales. However, to study strongly nonlinear phenomena and interactions, the machine learning methods need to be sufficiently flexible to account for such nonlinearities. This often requires many parameters to tune and a good understanding of the inner workings of an algorithm. On the other hand, in climate science domain knowledge is commonly used to preprocess the data before the actual analysis. For example, we might want to remove the seasonal cycle or interannual variability if we are interested in studying a phenomenon that is on a lower temporal scale. However, such preprocessings lead to loss of information about the interactions between the different scales. The open question is how to combine the domain knowledge from climate science and the capacity of machine learning for discovery in order to gain further understanding. This would most probably require an iterative process and feedback between climate scientists and data scientists. The last part of the talk will bring up the problem of interdisciplinarity and the growing interest of machine learning students to work on climate projects, especially climate change. How can we leverage this interest in order to advance the field of climate science?

18

The wrong dichotomy: Data science as supplementing, rather than displacing other methods

Dominik Fitze[1]

[1]*University of Bern*

Sometimes in philosophy and the wider public, there is an implicit or explicit assumption that data science and traditional research methods are opposed to each other. This sentiment can be traced back to at least Anderson's 2008 WIRED Op-ed, which has been discussed often in philosophy.

I contend that this discussion is based on wrong premises. In reality, machine learning and "traditional" methods often go hand in hand. Drawing on Reichenstein et al. (2019) and a short case study, I will argue that Machine Learning and theory-based modeling are likely to go hand in hand, and that such applications can already be observed in climate science. The picture that emerges is one where climate models can be enhanced by supplementing or even replacing some of its parts by machine learning approaches, while other parts will retain "traditional" methods, leading to a mix of data science and physics-based model parts.

Finally, I will relate those findings to similar research in biology (Canali 2016, López-Rubio & Ratti 2019) to show that machine learning appears to be another powerful tool in the toolbox of many scientists in different disciplines. The same should be expected for the future of climate science.

Interdisciplinarity and Scientific Practice

Friday, Parallel Session 1, 14:30 – 16:30

19

Challenges and Opportunities in the Publication of Global-scale, High-resolution Climate Model Outputs

Ionut Iosifescu Enescu[1], Gian-Kasper Plattner[1], Dirk Nikolaus Karger[1], David Hanimann[1], Dominik Haas-Artho[1], Martin Hägeli[1], Niklaus E. Zimmermann, Konrad Steffen[1,2,3]

[1]Swiss Federal Institute for Forest, Snow and Landscape WSL, [2]École Polytechnique Fédérale de Lausanne EPFL, [3]ETH Zurich

EnviDat is the environmental data portal of the Swiss Federal Research Institute WSL offering a wide range of support services for research data management and data publication [Iosifescu et al. 2018, 2019]. It actively implements the FAIR (Findability, Accessibility, Interoperability and Reusability) principles, offering formal publication of research data with proper citation information and Document Object Identifiers (DOIs).

The publication of global-scale, high-resolution climate model outputs poses major challenges for our portal in its current setup, but also opens up new exciting opportunities. Recently, the “climatologies at high resolution for the Earth’s land surface areas (CHELSA)” products are being made available through EnviDat. CHELSA provides free climate data at 1km resolution for various time periods as documented in Karger et al. (2017).

The first EnviDat challenge for the publication of such global-scale climate models’ outputs is their size. Terabytes of data are being produced for each CHELSA version and the size is expected to increase to petabytes. Moving towards the publication of Singularity containers and additional resources that would allow to reproduce the actual research data represents a possible, albeit limited solution.

The second EnviDat challenge is related to the versioning of such climate model outputs, as DOI-ed datasets can no longer be deleted and each version of the data must be kept, in principle, forever. Consequently, workflows and technical solutions will also be needed for minimizing DOI-ed data redundancy.

The third EnviDat challenge is related to the visualization of the model outputs directly from a publishing portal. The direct map-based visualization of climate model outputs in an embedded WebGIS platform are novel requirements for environmental data publication.

Finally, in the above challenges lie also major opportunities for the community and a portal such as EnviDat. Significantly improving the publication and on-demand visualization of future climate data model outputs is one of our aspirations to support the community.

References:

Iosifescu Enescu, I., Plattner, G. K., Pernas, L. E., Haas-Artho, D., Bischof, S., Lehning, M., & Steffen, K. (2018). The EnviDat concept for an institutional environmental data portal. Data Science Journal, 17, 28 (17 pp.). https://doi.org/10.5334/dsj-2018-028

Iosifescu Enescu, I., Plattner, G. K., Bont, L., Fraefel, M., Meile, R., Kramer, T., … Steffen, K. (2019). Open science, knowledge sharing and reproducibility as drivers for the adoption of FOSS4G in environmental research. In M. A. Brovelli & A. F. Marin (Eds.), International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences: Vol. XLII-4/W14. FOSS4G 2019 – Academic Track (pp. 107-110). https://doi.org/10.5194/isprs-archives-XLII-4-W14-107-2019

Karger, D. N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., … Kessler, M. (2017). Climatologies at high resolution for the earth's land surface areas. Scientific Data, 4, 170122 (20 pp.). https://doi.org/10.1038/sdata.2017.12220

20

Situated knowledge and climate services: miscellaneous scales and levels of interpretation, using physical observations and data science

Mélissande Machefer[1], Arthur Brun[2]

[1]Lobelia by isardSAT, [2]Paris Nanterre University

Climate science often studies large-scale spatio-temporal phenomena depicted by long data records, many of which are now available from Earth Observation satellites on a recurrent basis. Data-dependent models are particularly good at generalising from high dimensional, heterogeneous and multi-sources information. The exploitation of these observations with data science as opposed to physical modeling is game-changing for operationalising climate services combining various study scales. In this sense, the panel of addressable environmental science problems requires a deep understanding of the involved phenomena to ensure the usefulness of data-driven methodologies.

Defining the physical and usage boundaries of the problem must respectively occur in collaboration with experts of the field and beneficiaries of the project. Considering that each of these agents acquaints with different contextual values, adopting an inducting risk view [Parker W. 2019] can mitigate the avoidance of cascade of uncontrolled biases which can arise from different nodes in the development chain and interpretability of the results.

Operational choices with data science allowing the automatic portability of the models across transversal use cases enable a pioneering “bird’s view” for large-scale phenomena whilst this global framework should also answer the granularity needs of the local beneficiaries for validation. However, all involved actors (researchers, data scientists, software developers, end-users ..) exhibit a pre-existent space of knowledge (in the sense that each and every actor’s knowledge comes from a positional perspective), or, as D. Haraway puts it, a situated knowledge, influencing their own requirements of goodness of fit. This paradigm highlights the need of situating the frame of subjectivity of each agent.

To which extent can the large scale results predicted with data-dependant models support decisions with the risk of “known unknowns” and taking into account “situated knowns”?

How does this paradox challenge the relevance of on-demand services for a large community of heterogeneous end-users?

21

Multilingual Structured Climate Research Data in Wikidata - The Community Perspective

Cristina Sarasua[1], Daniel Mietchen[2]

[1]University of Zurich, [2]School of Data Science, University of Virginia

Empirical sciences experience a transformation enabled by a myriad of technological solutions that facilitate collecting, sharing and analyzing large- and small-scale research data. Citation networks can be mined, scientific workflows can be reproduced and extended, and data-driven search portals allow scientists to dive into a sea with millions of data sets. While technology is crucial, the success of this transformation heavily depends on social change and commitment. At the core of such a social response, the Open Science movement promotes values such as participation and collaboration. Tightly connected to Open Science, the Free Knowledge initiative advocated by Wikimedia has succeeded in bringing scientific output (and general human knowledge) closer to the global population through platforms like Wikipedia. Wikidata is a community-supported knowledge base, where thousands of volunteers enter, complete, link, monitor and correct data. Wikidata is connected to Wikipedia articles and images in Wikimedia Commons, and it can be queried as machine-readable Linked Data. In this presentation, we would like to showcase Wikidata's special features in terms of collaborative knowledge management. We will demonstrate how ranks and references allow Wikidata to portray a plural reality in which contradictory statements might have been published by different sources. We will also demonstrate the way federated queries can facilitate data comparison. Moreover, we will describe the process that editors follow to address schema and data quality management collectively, as well as human-bot cooperation. We will also talk about the possibility of transferring many of Wikidata's features to self-organized communities via Wikibase. Through concrete examples and descriptive statistics, we aim to show the benefits that a community-based data management cycle can provide to many disciplines, including the field of Climate Research.

22

Multilingual Structured Climate Research Data in Wikidata - The Data Perspective

Daniel Mietchen[1], Cristina Sarasua[2]

[1]School of Data Science, University of Virginia, [2]University of Zurich

Climate research — like research in general — takes place in a sociotechnical ecosystem that connects researchers, institutions, funders, databases, locations, publications, methodologies and related concepts with the objects of study and the natural and cultural worlds around them.

Mechanisms for describing concepts related to climate research are growing in breadth and depth, number and popularity. In parallel, more and more climate-related data — and particularly metadata — are being made available under open licenses, which facilitates discoverability, reproducibility and reuse, as well as data integration.

Wikidata is a community-curated open knowledge base in which concepts covered in any Wikipedia — and beyond — can be described in a structured and FAIR fashion that can be mapped to RDF and queried using SPARQL as well as various other means. Its community of over 20,000 monthly contributors oversees a corpus of currently over 80 million ‘items’ for concepts that are linked amongst each other, to external databases or to specific values via over 7000 'properties'. Items and properties have persistent unique identifiers, to which labels, descriptions and dedicated lexemes and their forms and senses can be attached in over 300 natural languages.

A range of open-source tools is available to interact with Wikidata — to enter information, curate and query it. In this presentation — available via https://github.com/Daniel-Mietchen/events/blob/master/data-science-in-climate-and-climate-impact-research.md — we will outline a range of tools that allow to explore Wikidata content through frontends tailored to specific communities. In particular, we will take a look at Scholia, which is available via https://tools.wmflabs.org/scholia/ and allows to generate and explore scholarly profiles of authors, institutions, funders and other parts of the research ecosystem, as well as of the world in which it is embedded, from geomorphological features to economic indicators and environmental policies, from natural ecosystems and disasters to biogeochemical cycles.

Modeling and Representation

Thursday, Parallel Session 2, 14:30 – 16:30

23

(How) can machine learning be used for predictive modeling of the Earth system?

Benjamin Stocker[1]

[1]*ETH Zürich*

We have arrived in an era of big environmental data. Yet, this massive increase in the wealth of data has not translated into a comparable progress in our ability to predict the state and functioning of the Earth system under future conditions. This unsolved prediction challenge is particularly pressing in terrestrial ecology and the science of terrestrial biogeochemical cycles, where it is underpinned by the fact that ecosystems are now subjected to conditions that lie well outside the domain (climate and CO2) in which they established and have evolved and for which we have observational data from the past.

Does this suggest that the promise of big data cannot materialize to solve challenges in environmental and Earth system science? Does this invalidate the contention that pure empirical approaches can make theory obsolete, given sufficient data (Anderson, 2008)? Or have we not yet found the right approaches to effectively translate information into understanding of the Earth system?

Using the experience from my own work, I will discuss examples of research challenges in terrestrial ecosystem modeling that are amenable to solutions relying on big data and machine learning (ML) and examples where such approaches are prone to failure. I argue that theory and physical constraints cannot be superseded, but should be embodied in data-driven methods. I argue that addressing the following questions will be fruitful for expanding the scope of ML methods in Earth system sciences:

What are characteristics of problems where ML holds particular promise?

How can we use the potency of ML for finding patterns between variables and use this information in a predictive framework (e.g., an Earth system model)?

How can we lower the bar for the implementation of data science methods in Earth system sciences?

24

Building a Seasonal Earth System Model Emulator for local temperature

Shruti Nath[1,2], Quentin Lejeune[1], Lea Beusch[2], Carl Friedrich Schleussner[1], Sonia Seneviratne[2]

[1]*Climate Analytics,* [2]*ETH Zurich*

Emulators are statistical devices that derive simplified relationships from otherwise complex climate models. Their unique ability to cheaply reproduce additional model realisations allows free exploration of model parameterizations, climate sensitivities as well as future climate scenarios. A recently developed Earth System Model (ESM) emulator, MESMER (Beusch et al. 2019), uses pattern scaling to provide spatially resolved yearly temperature

values from global mean temperature values. Through a novel innovation term, MESMER is furthermore able to represent internal climate variability, yielding a convincing imitation of the interannual variability of a multi-model initial condition ensemble. The work presented here extends MESMER's framework to have a seasonal downscaling module, so as to provide spatially resolved monthly temperature values from global mean temperature values. This is achieved by training a harmonic model on monthly ESM outputs in order to capture seasonal cycles and their evolution with changing temperature. Four ESM runs from CMIP5, RCP 8.5 are used with ensemble members split into training and test sets in a two to one fashion. Once the mean seasonal cycle is sufficiently emulated, an additional variability term is added to allow for some stochasticity and hence, prediction of extreme monthly climate events. This variability term maintains serial correlation between months through an Auto-Regressive feature. The biases of the seasonal downscaling module are evaluated in terms of spatial and temporal evolution. By scrutinising the change in biases along the temperature distribution, physical processes e.g. ice-albedo feedbacks poorly represented within the model are moreover contemplated. Outputs of this emulator are expected to provide impact assessment models with spatially and temporally resolved data for future scenario exploration.

25

Can Machines Learn Convection? The epistemic implications of machine-learning parameterizations in climate science

Suzanne Kawamleh[1]

[1]*Indiana University*

Scientists and decision makers rely on climate models for predictive insight concerning future climate change, particularly extreme events. However, many physical processes which are key to accurately predicting extreme events are indirectly represented in the model using physical parameterizations. Scientists are exploring and successfully using a machine learning approach to replace physically-based parameterizations with neural network parameterizations (NNPs).

I analyze the epistemic implications of the shift from physically-based to data-driven parameterization schemes. I argue that the training of a NNP on a previously-tuned high-resolution model (1) introduces an additional degree of freedom between the output and the observational data which contributes to the epistemic opacity of NNPs and (2) increases parametric uncertainty. Given the sensitivity of model projections (and their reliability) to parameterization schemes, I show that increased parametric uncertainty has negative implications for the accuracy and reliability of model projections. This is supported by the repeated failure of NNPs to successfully generalize outside the training data set.

The improved representation of model processes is one important way of improving model performance. The failure in NNP generalizability indicates that learning the quantitative relations that hold between climate variables is not adequate for representing the physical processes behind the output data. The direct or indirect representation of a process plays a crucial role in supporting model projectability or generalizability. The very representation of processes adds significant and irreplaceable value for the reliability of climate model predictions.

I conclude that NNPs cannot reap the predictive advantages of high resolution models while leaving out the key reason for their improved performance—the explicit and improved representation of sub-grid processes that govern model predictions. Rather, the adoption of

NNPs negatively impacts the way scientists interpret and manage the substantial uncertainties associated with climate model parameterizations, thus undermining the accuracy and reliability of model projections.

26

The role of machine learning in site-specific wind turbine power curve prediction

Sarah Barber[1]

[1]*University of Applied Sciences Rapperswil*

The accurate prediction of the power production of a wind turbine at a particular site is important in both the planning and operation phases; however, the standard power curve binning method is not specific to the atmospheric conditions at the site. Machine learning can be used for improving site-specific prediction of power curves, for example by applying regression trees to measured atmospheric conditions such as wind speed, turbulence intensity and shear factor.

This work starts by discussing some examples of applying regression trees to power predictions. The first example involves creating a set of 8,000 ten-minute long aero-hydro-servo-elastic simulations of the NREL 5MW reference wind turbine at a random combination of hub-height wind speeds, turbulence intensities and shear factors using cloud computing with the software ASHES. A regression tree with maximum depth of eight and optimised with Adaptive Boosting is trained using a random selection of half of the data. For a set of 50 random test cases, the Root Mean Square Error of the predicted power compared to the simulated power is found to be three times smaller than for the standard power curve method of binning. The second example applies the method to real measurement data from a wind farm in Brasil. The same method was applied, and the Root Mean Square Error of the predicted power compared to the simulated power is found to be 1.6 times smaller than for the standard power curve method of binning. The reduced accuracy compared to the first example is due to the non-uniform distribution and limited range of atmospheric conditions in the real wind farm.

Next, the suitability of applying machine learning to wind turbine power prediction in general is discussed, particularly considering if it can be used equally well for understanding as for prediction.

5. Social Activity

Thursday, 18.00 – 18.30, in the zoom meeting room of the keynote sessions.

The social activity will give you the opportunity to network with other participants from the workshop. To participate, just stay in the zoom meeting room after the keynote. We will then create break-out rooms with three to four participants and give you the chance to have informal talks for about 10 minutes before we regroup the participants. The goal is to give you the chance to meet other people from other institutes and potentially other disciplines such that you can connect later on and discuss areas of mutual interest.

6. Author Index

Here you can find all authors and the numbers of the abstracts to which they contributed. Abstract numbers are ascending in chronological order.

F

G

H

I

J

K

L

M

N

O

P

R

S

T

V

Z

Zimmermann, Niklaus E.	19	Zumwald, Marius	P3