Chapter 4
Data Analysis and Exploration with Computational Approaches

Viktoria Wichert, Laurens M. Bouwer, Nicola Abraham, Holger Brix, Ulrich Callies, Everardo González Ávalos, Lennart Christopher Marien, Volker Matthias, Patrick Michaelis, Daniela Rabe, Diana Rechid, Roland Ruhnke, Christian Scharun, Mahyar Valizadeh, Andrey Vlasenko, and Wolfgang zu Castell

Abstract Artificial intelligence and machine learning (ML) methods are increasingly applied in Earth system research, for improving data analysis, and model performance, and eventually system understanding. In the Digital Earth project, several ML approaches have been tested and applied, and are discussed in this chapter. These include data analysis using supervised learning and classification for detection of river levees and underwater ammunition; process estimation of methane emissions and for environmental health; point-to-space extrapolation of varying observed quantities; anomaly and event detection in spatial and temporal geoscientific datasets. We present the approaches and results, and finally, we provide some conclusions on the broad applications of these computational data exploration methods and approaches.

Keywords Machine learning · Artificial intelligence · Earth system · Data exploration

V. Wichert (✉) · N. Abraham · H. Brix · U. Callies · V. Matthias · A. Vlasenko
Helmholtz-Zentrum Hereon, Geesthacht, Germany
e-mail: viktoria.wichert@hereon.de

L. M. Bouwer · L. C. Marien · D. Rechid
Climate Service Center Germany (GERICS), Helmholtz-Zentrum Hereon, Hamburg, Germany

E. González Ávalos · P. Michaelis
GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel, Germany

D. Rabe · W. zu Castell
Helmholtz Centre Potsdam—GFZ German Research Centre for Geosciences, Potsdam, Germany

R. Ruhnke · C. Scharun
Karlsruhe Institute of Technology, Eggenstein-Leopoldshafen, Germany

M. Valizadeh · W. zu Castell
Helmholtz Zentrum München—German Research Center for Environmental Health, Neuherberg, Germany

L. M. Bouwer et al. (eds.), *Integrating Data Science and Earth Science*,
SpringerBriefs in Earth System Sciences,
https://doi.org/10.1007/978-3-030-99546-1_4

4.1 Introduction and Challenge

Computational data exploration and analysis can help to substantially improve modelling and understanding of Earth system processes. In this chapter, we provide an overview of the developments in the Digital Earth project that focus on employing such innovative techniques to improve our process understanding, to derive new insights from a variety of existing datasets and to make the investigation of complex processes more feasible. Diverse sub-disciplines in the Earth sciences are using computational methods to solve some of the major issues identified for the Earth science community. The issues for which computational applications have been developed in Digital Earth and that are presented in this chapter are as follows:

- Extracting relevant information and features using machine learning approaches: For various features, no labelled data collections exist, as these are too labour intensive to develop. Labels are important for supervised learning algorithms, for example, to classify specific observations using prior knowledge. Using sparse datasets and machine learning methods, alternative ways can be found to broaden data availability and derive new, crucial information from existing data. Here, examples are provided that map river levees in Germany, for which no consistent data were readily available for research before and for detecting underwater ammunition locations.
- Approximating complex processes with machine learning: Although models are successfully used to understand complex processes in the Earth system, for some applications the computational cost is too high to embed this information in a broader framework and to answer questions that are more challenging. In these cases, approximating these same processes through machine learning algorithms can be a means for scientists to tackle those problems. We present an approach to estimate methane and ethane concentrations through a Neural Network and give an example of how machine learning algorithms can be used to combine highly heterogeneous data to answer pressing questions related to climate change and health research.
- Point-to-space extrapolation: For many applications in the Earth sciences, single-point observations need to be inter- and extrapolated across space to arrive at consistent estimates of total matter fluxes. This chapter presents an example where point observations of methane emissions are analysed and processed in order to be consistent with global atmospheric emissions as observed/estimated in global databases. A second application provides insights into the functionalities of advanced approaches for point-to-space extrapolation.
- Anomaly and event detection across heterogeneous datasets: In some applications, process understanding can be improved considerably when data from diverse sources are combined through computational methods. Detection of events and anomalies is important in Earth systems for scientific and for practical applications. We present an approach that combines observational and model data to detect river plumes at sea at the end of a riverine flood event chain and tracks their spatial and temporal extent.

4.2 Object Recognition Using Machine Learning

4.2.1 Deep Learning Support for Identifying Uncharted Levees in Germany

Germany has a large network of levees to manage flood events. Unfortunately, data on the locations of these levees are not always directly available to the public or to researchers. Due to the importance of these levees in the analysis of flood events, approaches are needed to derive the levees' location and height from available information. However, such methods are not readily available, and neither are commonly accepted nor standardized approaches. Advances in computational methods, namely deep learning, and the release of a wide range of geodata to the public make it possible to find levees automatically and on a large scale.

In this research, we have started to develop such a framework and appropriate methods to delineate levee features from such data. As data sources, we combined aerial images and LIDAR-based digital elevation models. The raster format of this data is comparable to image data, and as such, we apply deep learning methods, which is providing state-of-the-art solutions for various computer vision problems. For example, in medical image analysis, deep learning models are used for cell classification and the tumour detection.

Our approach relies on semantic segmentation, a common task in deep learning. Semantic segmentation refers to the classification of individual pixels to different predefined classes. The output has the same raster shape as the input features. To train a semantic segmentation model, a mask with a classification of the input features is needed. In our case, these input features are the pixels from the aerial images and digital elevation model. We choose a common architecture for our semantic segmentation model, the U-Net (Ronneberger et al. 2015). A U-Net consists of blocks of convolution layers in combination with pooling (to reduce the data size by a factor of two) or upscaling (to increase the data size by a factor of two) layers. There is an equal number of pooling and upscaling blocks. The pooling blocks come first, followed by the upscaling blocks, transferring the information from input to output. Additionally, the output of the first pooling blocks is used as an input for the last upscaling block, and the output from the second pooling block is used as an input for the second to last upscaling block. This schema continues for the other blocks in the network.

To train our model, we used the publicly available data from North Rhine-Westphalia (NRW 2021), which contains the relevant information. As features, we use the LIDAR-based digital elevation model (1 m resolution in two by two-kilometre tiles) and the aerial photographs (0.1 m resolution in one by one-kilometre tiles). The levees are available as shapefiles for the crest of the levees. Several pre-processing steps are applied.

The first step is to rescale the aerial images to the resolution of the digital elevation model and to split the digital elevation model into one by one-kilometre tiles. Afterwards, we rescale the pixel values for all inputs to a range of zero to one. The

processing of the labels is more complex, as we need to create a mask from the line shapes. The lines themselves only cover a small subset of the pixels; therefore, we use the entire width of the levees, which is not given in the dataset. To derive the width of a levee, we take orthogonal sections to each line segment and look for the maxima in the second derivative of the digital elevation model along the sections. We use these maxima as the boundaries of the levees. In addition to the derived masks for levees, we also include a mask for the bodies of water which is derived from publicly available polygon shapes. The training itself is run on GPUs and includes common data augmentation techniques such as rotations.

The results must be in the same format as the original input, so we have to process the output of the neural network and extract the information. This post-processing consists of multiple steps. The first step is to assign contiguous areas of pixels to different groups. This is based on a threshold value as the neural network output is the probability of a pixel to be of one class. We also apply a maximum filter of size four to remove some noise which can be induced by applying the threshold. For each group of pixels, we then look for the highest points as we want to find the crest of the levee. These points should all fall in the same height range, a criterion we use to exclude unlikely levees. The next step is to merge groups of adjacent tiles to get the entire levee as one object. Additionally, we analyse the length of the crests and discard short sections. All points of the merged groups are then transferred to a line shape. The shape is simplified using standard tools to reduce the number of points specifying the shape. These output shapes can then be used for analyses. We additionally use the pre-processing method to create polygons for the entire levees (c.f. Fig. 4.1). Overall, the methodology detects a high percentage of the charted levees, where

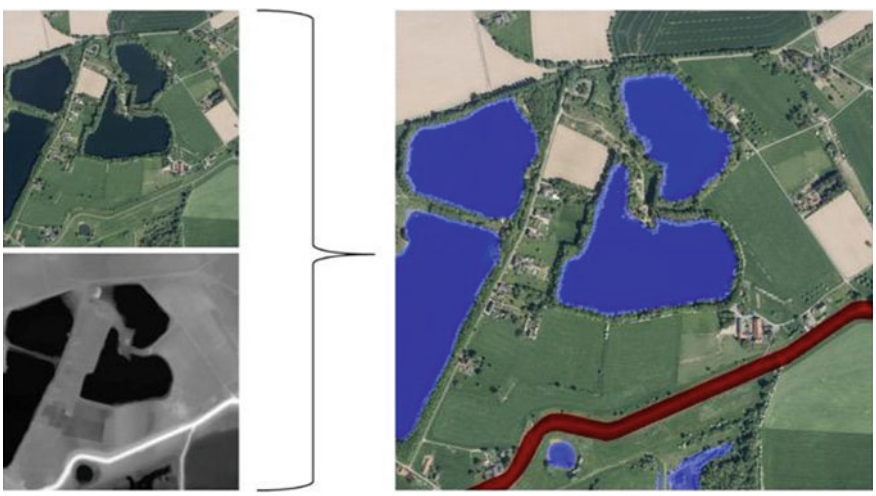

Fig. 4.1 Exemplary output of the deep learning model after post-processing. The input data (aerial image on *top left* and digital elevation model at the *bottom left*) together with the output, the aerial image with the predicted shapes as overlays

precision and recall can be balanced by adjusting the selection threshold. The next step is to evaluate the methodology and model using data from a different state, e.g. Saxony. In general, this approach is applicable to a wide range of problems in remote sensing. The fusion of different data sources is still uncommon in combination with deep learning models. Our use case highlights the benefit of such an approach.

4.2.2 Machine Learning Support for Automated Munition Detection in the Seabed

Both the North Sea and Baltic Sea have been used as dumping grounds for munitions, especially after the Second World War. Nowadays, various infrastructure projects are planned and built in these waters, including offshore wind farms and pipelines. Before construction can begin, the area must be cleared of munitions. Multi-beam echo sounders, side-scan echo sounders, magnetometers and sub-bottom profilers are used to explore the construction sites. It is, however, very time consuming to analyse the data generated by the instruments. We propose the use of machine learning to detect munitions in the data.

For our analysis, we rely on the multi-beam echo sounder. The echo sounder dataset provides the water depth and the backscatter, giving information on the composition of the seafloor, at a horizontal resolution of 25 cm (where depths are below 20 m). The integration of other data sources, e.g. magnetometer data, into one model is possible, but comes with a number of challenges. The most important one is getting an accurate spatial alignment of the datasets. A mismatch of even one metre can cause many objects not to overlap in the combined dataset.

The machine learning method we applied is deep learning. As with the previous application for levee detection, we face a semantic segmentation problem. The model class we use to approach this problem is a U-Net (Ronneberger et al. 2015). After testing various model configurations with depth (number of layers) and width (number of filters in a layer), we can conclude that the problem can be solved with relatively small models (with respect to both the depth and width). These models can be trained on the CPU within less than an hour.

To create a training dataset for our machine learning model, a few steps are necessary. First, we need a labelled dataset to be able to train a model. The second step is the preparation of the labels for the model type. In our case for multi-beam echosounder data, we need to create a mask for the map, based on available point labels. We achieved this by labelling an area around the point label as targets. This way we sometimes falsely label data as targets, but also account for imprecisely placed labels. To facilitate the training of a model, we scale and standardize the data in terms of the range of values. Here, we keep track of the exact steps and parameters used to be able to apply the model and processing steps to other datasets in the future. Additionally, we use standard machine learning practices like train-test splits and augmentation techniques like rotations and mirroring.

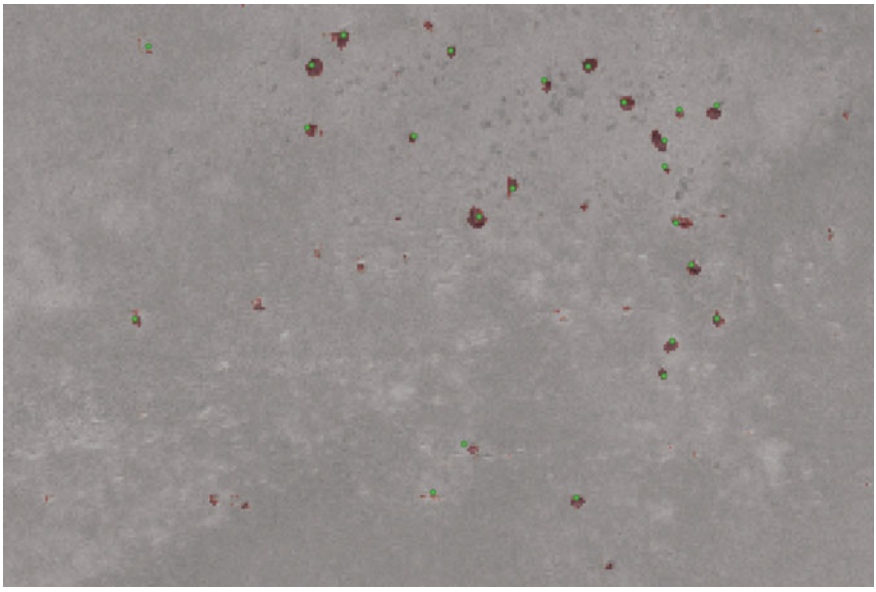

Fig. 4.2 Prediction of the neural network on top of the backscatter. *Red areas* indicate a high predicted probability for munitions, and clear areas indicate a low predicted probability. The *green dots* are the original labels

The results of the model (Fig. 4.2) look very promising as we can detect most of the labelled objects (>95%). However, we get several false positives, at least according to the labels, which might not be complete. The second important consideration is that the labels are not validated and might be false. Therefore, the number of false positives might be even higher.

We transferred this approach to a sub-bottom profiler. In both cases, we have two-dimensional data. The difference is that the data in this case are not in latitude and longitude direction but along a transect in latitude and longitude with the depth as the second dimension. One observation is that the models must be more complex to get good results. An important difference is in the training process where rotations are not a viable augmentation technique because objects create a distinct bell-shaped curve in the sub-bottom profiler images. The training can be done on a CPU as well. Overall, the issue of false positives persists while most labelled objects are found.

4.3 Approximating Complex Processes with Machine Learning

4.3.1 Estimation of Methane and Ethane Concentrations in the Atmosphere Over Europe by Means of a Neural Network

Methane is an important atmospheric greenhouse gas that has a substantial impact on climate and air quality (Van Dingenen et al. 2018). Chemical transport models (CTMs) are the most widely used tools allowing us to predict its concentration in air and its possible effects on the environment. A typical CTM accepts emissions and meteorological data as input data and calculates the concentration changes of atmospheric methane (and other gases) in time. Although CTMs have been continuously improved, they require a significant amount of computational resources (CPU time, RAM and disc space). Neural networks may become a cheaper alternative to CTMs in terms of these resources. The idea of a neural network (NN) is to fit a combination of simple mathematical functions (neurons or activation functions) so that for a given set of predictors and predictands, a NN, receiving predictors as inputs, estimates outputs that have minimal difference with the corresponding observed predictands. After training, the NN should be able to predict unknown output based on any given input. To verify its predictive skill, the NN is tested on an independent set of known inputs and outputs that were not employed in training.

To estimate methane concentrations, we developed two neural networks. Note that 19% of atmospheric methane is associated with fossil fuel production, mainly related to oil and gas mining (Van Dingenen et al. 2018); more than half of it leaks directly from the gas or oil fields. To detect these leakages from the offshore fields, we developed the first NN that estimates local anomalies in methane concentration directly from measurements near the potential (or known) natural gas source. We built this NN using Keras/Tensorflow package (Abadi et al. 2016). It consists of three dense layers with a hyperbolic tangent activation function, eight inputs (latitude, longitude, the temperature at two-metre height, time, humidity, latitudinal and longitudinal wind components at ten metres, sea surface temperature) and one output (methane). The NN was trained on the cruise measurements POS-526 (Greinert and Schoening 2019) that took place from 07.23.2018 to 11.08.2018 on a route: Bergen (Norway)—Dogger Bank (Netherlands)—Hirtshals (Denmark)—Tisler (Norway). The cruise data contain all input and output variables. Note that having geographical coordinates as inputs, the NN "learned" during the training the positions (and the corresponding emissions) of wells and oil fields by fitting its estimates to the methane concentration at different locations in the training data. After training, this NN, installed on a laptop, can estimate methane concentration at the current location from the current physical parameters of the surrounding atmosphere. If this estimate does not match the measured concentration, one may suspect to have detected an anomaly possibly associated with new oil or gas fields, or substantial changes in the known ones.

Note that measured methane anomaly may originate from other sources. To exclude the impact from other sources, we developed a second NN, which estimates daily mean ethane concentration anomalies in the atmosphere. Natural gas contains up to several per cent of ethane, giving 62% of atmospheric ethane (Franco et al. 2016). The concentration ratio of ethane/methane is unique for each oil or gas field and constant in time, serving as a kind of fingerprint (Visschedijk et al. 2018). Since the methane/ethane ratios near oil or gas fields are known (Yacovitch et al. 2020), we can estimate the fraction of atmospheric methane leaking from the gas or oil fields just from ethane. The second NN was developed on the basis of the network described in (Vlasenko et al. 2021). We trained and tested the second NN on the ethane anomalies estimated from the Consortium Multiscale Air Quality Model (CMAQ) (Appel et al. 2013) in the European domain for the period 1979–2009 using the same emission data for the year 2012 (Bieser et al. 2011) for the entire 30-year period. We define ethane anomalies as the deviation of the current value from its climatological mean. For the second NN, we again used the Keras/Tensorflow package (Abadi et al. 2016), choosing a NN that consists of one recurrent layer followed by two dense layers. It accepts wind anomalies in the European domain and estimates the corresponding ethane anomaly in the same area. All layers have hyperbolic tangent activation functions.

To train and test the NN, we split the data for both NNs as follows. The training set for the first NN contains 90% of samples, which is 14,400 inputs and outputs, randomly picked up from the POS-526 cruise measurements (Greinert and Schoening 2019). The remaining 10% of data, i.e. 1600 samples, were used for testing. For the second NN, we took estimated wind and ethane anomalies from 1979 to 2006 for training and anomalies from 2007 to 2009 for testing. Developing the second NN, we found that the inter-seasonal variability in the data resulted in errors. To minimize these errors, we created and trained the NN for each season separately.

To evaluate the accuracy of both NNs, we used the R^2 measure that shows how much of the observed data variability is explained by the model (which is the NN in our case). Note that for the second NN, anomalies obtained from CMAQ estimates play the role of observations.

The estimates of methane and ethane obtained from the first and the second NN are shown in Figs. 4.3 and 4.4, respectively. Note that the estimates of the first network (blue line) and the measured methane concentration (red line) almost coincide. As a result, R^2 for the first network equals 0.91. The R^2 for the second NN equals 0.565, 0.48 and 0.57 for summer, spring and winter, respectively. Although the second NN has lower R^2 than the first NN, it reconstructs the ethane anomaly patterns' main features. This can be seen in comparison with the summer mean ethane anomalies estimated by the NN (Fig. 4.4, left panel) and CMAQ (Fig. 4.4, right panel). Note that except for small details, the NN succeeds in reconstructing the pattern of the CMAQ simulations. Other deteriorations of R^2 are caused mainly by a slight underestimation of the anomaly amplitude. Relying on these results, we conclude that the first and the second neural networks predict the corresponding methane and ethane concentration anomalies with a high degree of accuracy and can be used as a smart monitoring tool during the researcher campaigns. Combining these neural networks into one

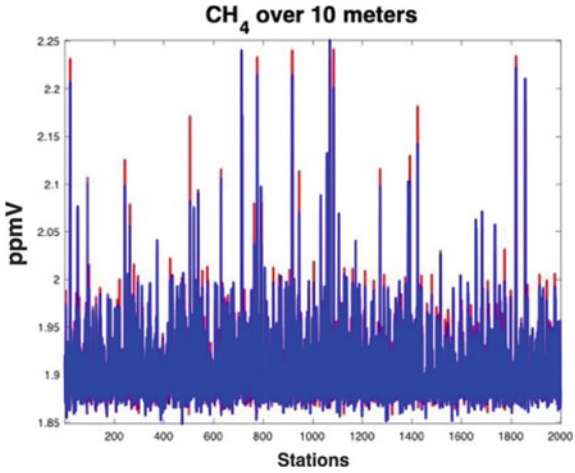

Fig. 4.3 Estimated (*blue*) and observed (*red*) methane concentrations, corresponding to the measurements in cruise POS-526 in the North Sea

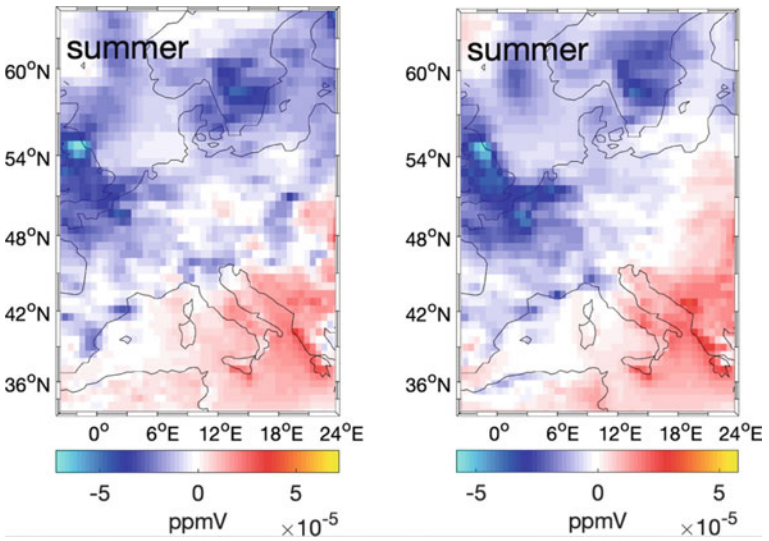

Fig. 4.4 Mean summer ethane concentration anomalies estimated with CMAQ (*left*) and NN (*right*)

predictive system, enabling the more accurate determination of the source of methane leaks is the next step in their development.

4.3.2 *Fusing Highly Heterogeneous Data to Facilitate Supervised Machine Learning in the Context of Health and Climate Research*

Heat waves can significantly affect human health. Examples include increased transmission of vector-borne diseases as well as increased susceptibility to metabolic conditions and higher mortality during episodes of severe heat. It is therefore paramount to investigate climate change in terms of potential-related health outcomes. To that end, the focus of our research is on temperature extremes, such as encountered during heat waves, and myocardial infarction (MI). Data from the region of Augsburg, Germany, are used as a case study. Epidemiological studies have shown that temperature extremes may indeed lead to an increased occurrence of MI (e.g. Chen et al. 2019). In future, frequency, duration and intensity of heat waves are expected to increase due to anthropogenic climate change, even at levels limited to 1.5° or 2° global warming (Sieck et al. 2020). Therefore, assessing health risks in the context of climate change is important for supporting more climate-resilient societies, public planning and adaptation strategies for human health.

Machine learning (ML) is a powerful tool for investigating complex and unknown relationships between environmental conditions and their adverse impacts and has already been applied in other fields (e.g. Wagenaar et al. 2017). ML is a data-driven approach, and meaningful results depend on consistent and high-quality data. To investigate climate change and MI, not only climate and meteorological data are required, but confounding effects of other well-known risk factors must also be accounted for by additional environmental, demographic, behavioural and socio-economic data. This makes this research challenging, as availability, provenance, and detail or resolution (temporal and spatial) of the data are highly variable. Here, we present a dedicated approach, designed to fuse such heterogeneous data into a consistent input dataset for ML algorithms.

The main pillar of our data-driven approach is the KORA cohort study (Holle et al. 2005) and the MI Registry in the Augsburg region of Bavaria, Germany. This dataset comprises detailed information on MI occurrence and underlying health conditions. Based on the registry data, a daily time series of MI incidence in Augsburg and two adjacent districts can be derived. This provides the target values for the training data for the ML algorithms.

To learn about the association of MI and the diverse risk factors such as exposure to heat, demographic structure, confounding health factors and air quality, these must be provided as predictors for the algorithms. Currently, weather and climate data (temperature observations and climate projections from the EURO-CORDEX initiative see Jacob et al. 2014); air pollution data (e.g. PM_{10}, PM2.5, nitrogen oxides and ozone; from regional environmental governments, such as BLFU, 2021); distribution of green spaces based on NDVI; demographic characteristics (age and sex, from regional statistical agencies, such as BLFS 2021); pre-existing illnesses (e.g. diabetes and obesity as recorded in the MI registry data); and socio-economic data (e.g. household income, education) are planned to be used within the project.

The raw data for these predictors are extremely heterogeneous for many reasons. First, the data come from different providers and is presented in various file formats, some of which are proprietary. Second, the representation of data can differ as well. Some data are of gridded/raster type (e.g. NDVI), some are point data (weather observations, air quality) and some are time series data, aggregated at the district level (e.g. demographic data). Third, the spatiotemporal scales (e.g. regional vs. local, coverage in time) and resolutions differ substantially.

The MI registry data are given in the form of individual cases with information on the date of the MI, the district where it happened and additional information on patient health. In a first step, a daily time series of MI incidence, aggregated at district level, is derived. The procedure is designed to produce compatible time series from the predictor data while addressing the three dimensions of heterogeneity outlined above. Afterwards, ML algorithms can readily be used to learn the relationship between incidence and the various predictors.

Figure 4.5 shows a schematic of the fusion procedure. Different predictors enter the pipeline and are subjected to a number of processing steps. The result is district-aggregated time series that can readily be used together with the MI data as input to ML algorithms. Depending on the nature of the predictor, only a subset of the processing steps may apply. First, the raw data are converted to a common format (csv). In many cases, the spatial scale is much larger than the region of interest. For instance, the NDVI data are global. The second step is therefore to reduce the data to the region of interest (ROI), namely Augsburg and surroundings.

Point sources, such as station data, are converted to a 1 km grid using a Kriging method. To account for the different resolutions in time interpolation to a common target frequency is conducted. The frequency is based on the highest resolution supported by the MI registry which is daily.

Finally, the data are aggregated to the district level to arrive at a daily time series for each of the predictors. Together with the ground truth, it can be used with all standard ML algorithms for time series prediction. The procedure has been implemented with the Python programming language. Figure 4.5 lists the packages used to carry out the processing steps.

The next step is to apply supervised ML techniques (e.g. Decision Trees, ANN) to the fused data to regress the incidence of MI based on the prepared environmental, socio-economic and climatic predictors. From this, we expect to gain first insights into the importance of heat stressors relative to other risk factors (Marien et al. 2022).

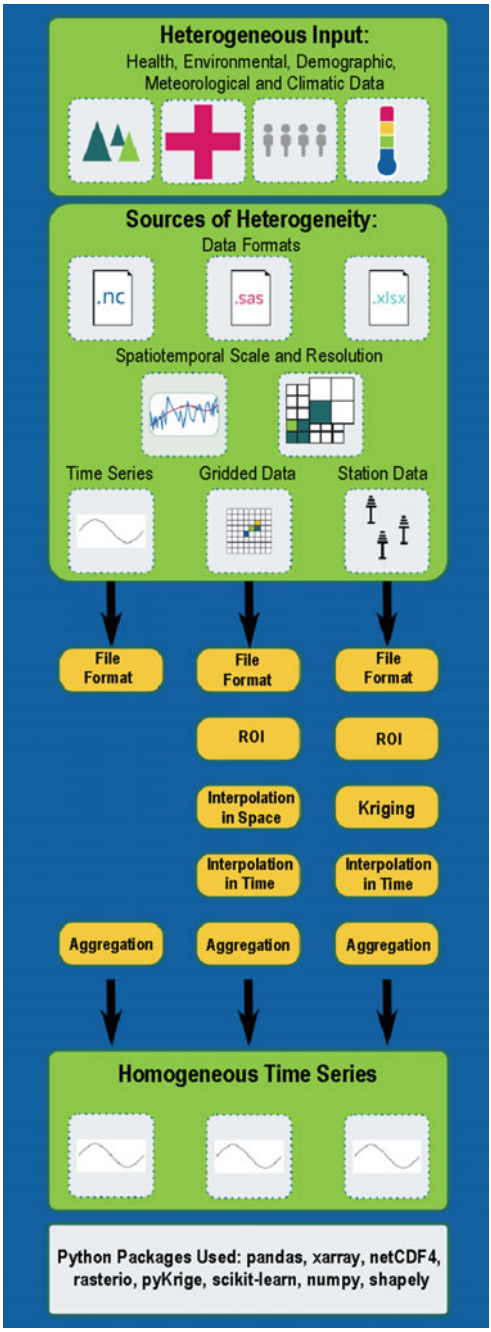

Fig. 4.5 Schematic of the fusion process chain

4.4 Point-To-Space Extrapolation

4.4.1 Estimation of Missing Methane Emissions from Offshore Platforms by a Data Analysis of the Emission Database for Global Atmospheric Research (EDGAR)

Disused and active offshore platforms can emit methane, the amount being difficult to quantify. In addition, explorations of the sea floor in the North Sea showed a release of methane near the boreholes of both, oil and gas-producing platforms. The basis of this study is the established Emission Database for Global Atmospheric Research (EDGAR) (Janssens-Maenhout et al. 2019). While methane emission fluxes in the EDGAR inventory and platform locations are matching for most of the oil platforms, almost all of the gas platform sources are missing in the database. We develop a method for estimating the missing sources based on the EDGAR emission inventory.

EDGAR is an inventory from the EC-JRC and Netherlands' Environmental Assessment Agency (Saunois et al. 2016). National reports of greenhouse gas emissions are the basis for emission inventories like EDGAR which is used as emission input for the simulations in this work. It covers sector- and country-specific time series of the period 1970–2012 with monthly resolution and a global spatial resolution of $0.1° \times 0.1°$ providing CH_4, CO_2, CO, SO_2, NO_x, C_2H_6, C_3H_8 and many other species. Different source sectors in EDGAR are defined using the IPCC 1996 guidelines (Janssens-Maenhout et al. 2019). When calculating the sector-specific emissions, a differentiation of emission processes improves and refines the estimates of EDGAR. Therefore, technology-specific emission factors, end-of-pipe abatement measurements, a modelling based on latest scientific knowledge, available global statistics and IPCC-recommended data are used. The emissions are then distributed on maps via proxy datasets based on national spatial data containing information about population density, the road network, waterways, aviation and shipping trajectories (Janssens-Maenhout et al. 2012). A global $0.1° \times 0.1°$ grid is used on which the emissions are assigned to, either as a single-point source (e.g. oil or gas platforms), distributed over a line source (e.g. shiptracks) or over an area source (e.g. agricultural fields) always depending on the source sectors and subsectors. For this work, methane point source emissions from EDGAR are of a high importance. These one-dimensional sources are allocated to a single grid cell of the $0.1° \times 0.1°$ grid with the average of all points that fall into the same cell (Janssens-Maenhout et al. 2019). We are aiming to replace the gridded emissions from EDGAR with point source in our atmospheric model by extrapolating them from point to space to adjust missing emissions within EDGAR and improve the spatial accuracy of the current dataset.

For this study, the global atmospheric model ICON (ICOsahedral Nonhydrostatic model) was used with EDGAR data as input for emissions. ICON is a joint development of the German Weather Service (DWD) and the Max Planck Institute for

Meteorology (MPI-M). Due to its dynamical core, which is based on the nonhydrostatic formulation of the vertical momentum equation, simulations with a high horizontal resolution up to grid spacings of a few hundreds of metres are possible. ART (Aerosols and Reactive Trace Gases) is an online-coupled model extension for ICON that includes chemical gases and aerosols. One aim of the model is the simulation of interactions between the trace substances and the state of the atmosphere by coupling the spatiotemporal evolution of tracers with atmospheric processes (Schröter et al. 2018). The point source module in ART takes the prescribed emission fluxes as point sources of substances and adds them to new or existing chemical tracers by distributing the one-dimensional fluxes to the area of the corresponding triangular grid cell in ICON. First of all, the emission factor is calculated through the source strength of the emissions, the area of the grid cell and the model time step as shown in Eq. (1) (Prill et al. 2019, p. 77).

$$\text{emiss\_fct} = \frac{\text{source\_strength}}{\text{cell\_area}} \cdot \text{dtime} \quad \left[\frac{\text{kg}}{m^2}\right] \tag{1}$$

As a next step, the above-calculated emission factor is added to the actual tracer value of the grid. Equation (2) shows how emiss_fct is multiplied with a height factor. Also the density of air ρ in kg/m^3 and the height of the corresponding ICON layer dz in metres come to play. In our case, all the point sources are on the lowest model level with a height of 20 m.

$$\text{tracer} = \text{tracer} + \frac{h \cdot \text{emiss\_fct}}{(\rho \cdot \text{dz})} \quad \left[\frac{\text{kg}}{\text{kg}}\right] \tag{2}$$

The ICON-ART sensitivity simulations of this study aimed to investigate the differences between simulations with gridded emissions from EDGAR and point sources of ART. Therefore, the methane emissions in the North Sea Region that are contained in EDGAR were distributed to all 956 point sources representing the offshore platforms in this area. Figure 4.6 displays the procedure of how EDGAR gridded emissions are replaced and adjusted by point sources.

If we compare simulations with gridded EDGAR and simulation with point sources, it is remarkable that both fit quite well over the whole year. This can be seen as a proof of concept that the point source module of ART adjusts the platform emissions of EDGAR in a satisfying way. The maximum of the absolute difference over the year on a global scale is 4.755 ppbv and the difference of the annual mean is -0.342 ppbv, showing that the point source module slightly overestimates the gridded emissions with a difference of less than 0.02%. These results show that the point source module in ICON-ART can model methane emissions as well as conventional gridded input data with the advantage that the spatial accuracy of this point-to-space method is significantly better.

With the point source module of ICON-ART being a successful tool for point-to-space extrapolation, we are now able to include the missing platforms into the

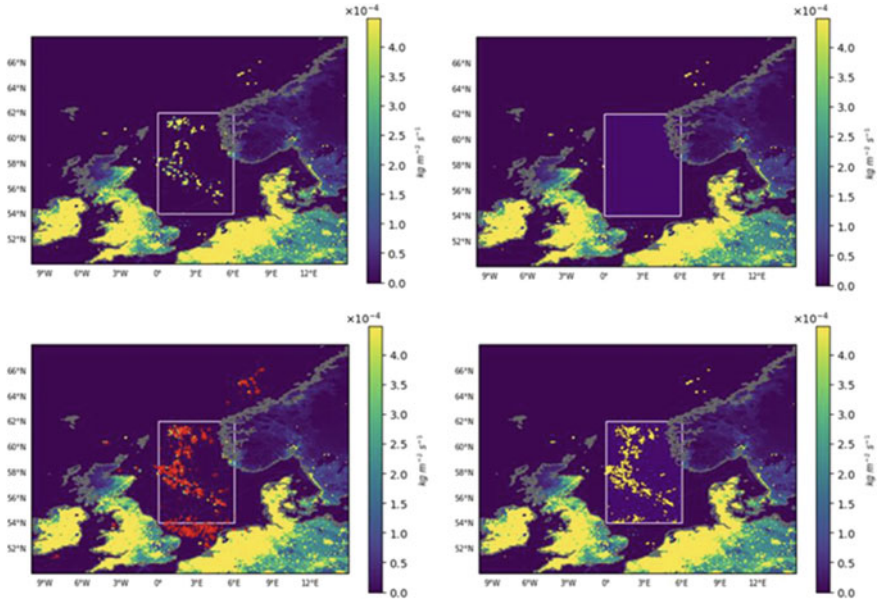

Fig. 4.6 Illustrating the procedure of how point sourceemissions adjust EDGAR gridded emission. The methane emission fluxes of original EDGAR (*upper left*) are removed within the North Sea Region (*white box*) (*upper right*). From all the locations of the platforms (*lower left*), the ones inside the North Sea Region are chosen and the EDGAR emission are equally distributed to them as point sources (*lower right*)

EDGAR dataset and access their influence on the methane distribution on regional (North Sea) and global scales.

4.4.2 Point-to-Space Methods

The aim of this subproject within the Digital Earth project is to develop a data-driven machine learning (ML) and artificial intelligence (AI) method to advance currently available datasets and maps, improve the resolution of observational datasets and interpolate values for locations where no observations are available, infer patterns from simulated data and improve estimations by combining the simulations with measurements.

Data sources at scattered sites (in situ measurements) and gridded data (satellite data, output of numerical models) are combined to construct a higher-resolution map, or predict values at ungauged sites (interpolation, "downscaling"), or to develop simulations for unobserved scenarios—locations and instance of time.

The project is designed as a close interdisciplinary collaboration between data science and Earth scientists. The fundamental concept of a point-to-space problem is

to find a solution in agreement with multiple data sources and to build a geographic dataset (or map) based on them. The outcome is a more complete dataset derived by extrapolating into space and time that can be used for further analyses.

This method can be used especially well to extend the sparsely scattered spatial and spatiotemporal maps by the application of ML methods instead of physical modelling approaches (Peng Xie et al. 2020; Amato et al. 2020, Volfová et al. 2012).

To build a procedure which can use multiple covariates from point or gridded datasets as inputs and find an estimator for the outcome variable, the following steps are required. First, in the pre-processing step, the data homogeneity and its ubiquity is checked.

Similar to most methods dealing with observation data, the raw data can be extremely heterogeneous and in various file formats, and adhere to different data standards. They can also differ in spatial and temporal resolution; hence, choosing a proper target grid resolution can be of utmost importance. In this step, all the data are projected into the same format and same target grid since the sources are so diverse. Often the first major issue is to pre-process and also check the observational data for errors and outliers. However, this problem is becoming simpler to solve with standardization of formats and metadata specification.

Normalization as second step in the pre-processing is another important factor, since the covariates can be of a different order, and hence, error propagation can be relevant to their scale (Singh et al. 2020). Determining the proper way to normalize is important, since it will have a direct effect on the algorithm as well as the results. In this step, all the variables, including dependent and independent variables, are normalized and hence prepared for the regression algorithms.

Given the nature of the problem, the respective Earth science expert needs to be consulted especially for the selection of the appropriate covariates. They need to be selected in a way that if the information is not available directly then it can be inferred from another variable or proxy. Moreover, the original point-to-space problem might extend similarly to other point-to-space subproblems for the selected covariates since they might not be available in the higher resolution of the target grid either.

A predictive ML regression model for these subsets of the problem can be trained and used to predict each of the covariates. For each of these covariates, a different ML method can be applied based on its characteristics and physics of the phenomena/covariate and the required precision. Consequently, after solving these subsets, the selected covariates are approximated on the target grid. The trained ML's prediction method can be applied, and this means that all the necessary requirements for the final step are now fulfilled.

In the final step, a more sophisticated ML regression method is applied for the target variable by combining the previously estimated covariates. Here, methods such as co-kriging to find optimal solutions for the outcome variable and covariates simultaneously can prove useful.

Similar to all ML methods, data splitting methods are employed and a percentage of the data is chosen as training data, with the remaining data used for testing and validation.

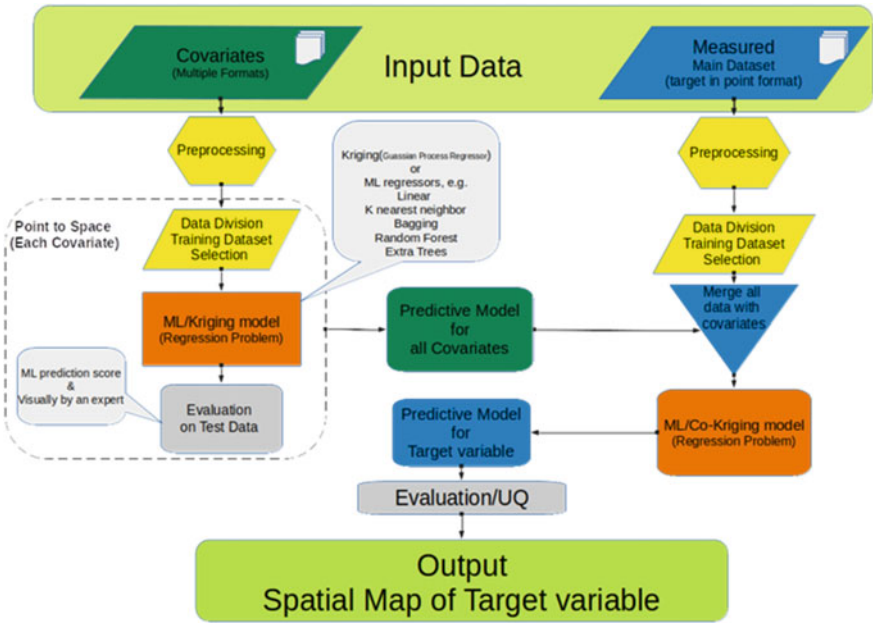

Fig. 4.7 Process chain schematic

Shown in Fig. 4.7 is a simplistic schematic of the suggested procedure. Any typical regressor can be easily implemented and used here, even a modified kriging method. Examples of the implemented and evaluated regression methods are linear, K-nearest neighbours, bagging trees, extra trees, random forest and simple ordinary kriging. Tree methods have been shown to have the advantage of high accuracy of the prediction for interpolations in the point-to-space problem. (Wessel et al. 2018).

To apply the same procedure to other similar problems, the following considerations need to be contemplated.

As with all ML regression problems, a regressor can lead to overfitting and some mitigation needs to be considered. One way is to employ different data splitting schemes and to check the number of features selected.

In co-kriging, similar to the normal kriging method, a kernel function is used based on a covariance matrix. Co-kriging is a conditional random field generator which superposes a method based on kriging with a multivariate Gaussian method building on a covariance matrix (Volfová et al. 2012). Determining the covariance function for co-kriging is not easy, since the methods with dependent covariates are so sensitive to the input and the stability of the system depends on the condition number of the prior (Putter et al. 2001; Ababou et al. 1994). Applying a co-kriging method enables a flexible feature advantage and makes it trivial to utilize any other data or information on new covariates/proxies which are available later, and as a result, the accuracy of the method can always be improved in this way.

Finally, uncertainty in data and methods should be evaluated to ensure that the method is based on a proper stable solution. Again, the evaluation of results using expert knowledge requires deep understanding of the problem, since data can be visually appealing but extremely erroneous.

This research is an example of a successful new collaboration between multiple centres involved in Digital Earth. It has been shown that it is indispensable to bring together knowledge from different fields and applications for developing successful applications.

4.5 Anomaly and Event Detection

In the Digital Earth Flood Showcase, we strive to understand hydrological extreme events across disciplines, from river basins to the sea. Comprehending the complex and often time-delayed chain of events surrounding a flood is a compartment-spanning task in the Earth sciences. Here, we feature an example from the showcase that illustrates how anomaly detection and investigation can be supported by suitable data exploration and analysis methods.

4.5.1 Computational Methods for Investigating the Impacts of the Elbe Flood 2013 on the German Bight

At the end of a hydrological extreme event chain, river water is discharged into the ocean, transporting matter and thereby, unusual amounts of nutrients and pollutants into the coastal system. To investigate the impacts of riverine floods on the marine environment, several steps need to be undertaken: the river plume needs to be detected and its spatial extent and development in time need to be determined to identify the study region and time interval. The procedure relies on the combination of heterogeneous data sources, such as in situ measurements, matching model data and additional satellite data for a wider spatial coverage. This article features three computational methods that are crucial for successfully identifying and investigating the processes in and around a river plume after a riverine flood event: automatic anomaly detection, producing customized model trajectories and generating time series of productivity. The complete workflow will be described in more detail in Sect. 5.3.3. under “The River Plume Workflow”.

Automatic Anomaly Detection

An enhanced feature of the River Plume Workflow is the automatic detection of parameter anomalies in both in situ FerryBox measurements during summer operations and year-round satellite data.

Three separate definitions of anomalies caused by a flood event originating from a freshwater river mouth are considered. They include: increased chlorophyll levels;

decreased salinity; and, particularly during the winter where the river water may be cooler than the sea, a decrease in sea surface temperature.

As an alternative to manually searching for anomalies in the River Plume Workflow's interactive map (see Sect. 3.2), users can select years of interest to undergo a Gaussian regression-based statistical analysis (Pedregosa et al. 2011), which provides the user with a list of recommended dates of interest. Gaussian processes (GPs) are powerful and flexible models for modelling time series data, which makes them a practical option for anomaly detection. Here, a Gaussian regression analysis is performed to generate a posterior probability distribution based on the daily parameters over the selected time period. This method requires a specified prior distribution, with the prior's covariance given by a kernel, in this case, a generalized Matern kernel with an amplitude factor and an observation noise component. The model uses high smoothing capabilities for more efficient anomaly detection. Also generated are posterior standard deviations that create a 95% confidence region around the posterior distribution. Anomalies are recorded when the measured data and its associated uncertainties fall completely outside the confidence region, either above or below depending on which parameter is being considered. Figure 4.8 gives an example of the outputted results for chlorophyll during 2013.

Future development would involve near real-time automatization to detect anomalies as they occur based on previous established annual patterns. This would involve taking an average of the posterior distributions over a number of years and detecting an anomaly in near real-time data if it deviates from this average.

Produce Customized Model Trajectories

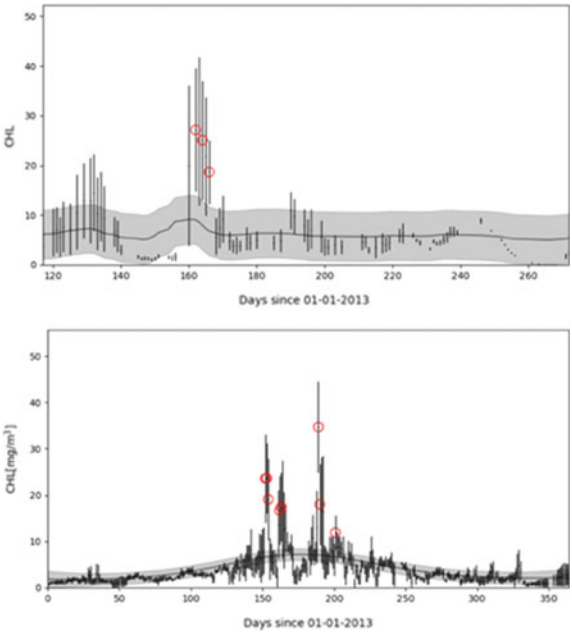

Fig. 4.8 Daily chlorophyll measurements with one region of interest (RoI) for in situ FerryBox (*upper*) and satellite (*lower*) data. The continuous line represents the Gaussian regression generated distribution, while the *shaded grey* band is the confidence region. The *red circles* indicate detected anomalies

A crucial component of the River Plume Workflow is the computation of model trajectories for observational FerryBox data to not only detect the river plume, but also determine its spatial and temporal extent in the ocean.

FerryBox transects provide, often regular, observational data on the study region. The data are made available in near real time via the COSYNA (Coastal Observing System for Northern and Arctic Seas) data portal (Baschek et al. 2017; Breitbach et al. 2016). The PELETS-2D code (Callies et al. 2021) is used to compute model trajectories for each FerryBox transect. These model trajectories consist of the measured water bodies' positions up to ten days before and after the actual measurement as simulated by the numeric model. The simulated positions are then recombined into synoptic maps, featuring time shifted positions of all observations at one point in time. While the model trajectories are useful to investigate an anomaly's origin, the synoptic maps show the spatial and temporal extent of the river plume on its presumed path across the ocean.

The simulations and their combination in synoptic plots are too computationally expensive to be done in real time. Therefore, they are currently only produced for specific events. However, work is currently underway to optimize the code for operational use and make the model data available in near real time.

Generate Time Series of Productivity

Combining observational data with specifically produced model trajectories allows users to determine the spatial and temporal extent of a river plume in the ocean, but does not help with understanding the processes that happen inside these waterbodies. For that reason, another method of the River Plume Workflow focuses on blending the spatial information of the model trajectories with the parameter information from satellite data. We use integrated satellite datasets taken from the Copernicus Marine Environment Monitoring System (CMEMS 2021). The datasets give daily average chlorophyll, salinity and sea surface temperature measurements after full processing, which includes the reconstruction of cloud-covered areas. The chlorophyll datasets have a 1 km^2 tiled resolution, while salinity and sea surface temperature have a 7×11.6 km^2 tiled resolution.

Our method automatically extracts parameter values from satellite data for the locations on a selected waterbodies' modelled trajectory, thus producing a time series of values along the modelled pathway of the waterbody. As an example, this method enables researchers to produce chlorophyll time series for waterbodies associated with the river plume and therefore to investigate chlorophyll degradation rates inside the river plume.

The methods described here were implemented into the River Plume Workflow, a scientific workflow prototype (see Sect. 5.3.3), and tested for the Elbe flood event from 2013. A ferry equipped with a FerryBox regularly covers the Büsum-Heligoland line during the period from April to October. These transects are highly relevant for our example as they cover a region close to the Elbe outflow. Using the anomaly detection algorithm, several instances where the Elbe river plume potentially crossed the FerryBox transect were determined, such as on 23 June 2013, where an anomaly in salinity and temperature was visible (see Fig. 4.9). Simulated trajectories of the relevant water bodies across the North Sea point to the Elbe River as the anomaly's

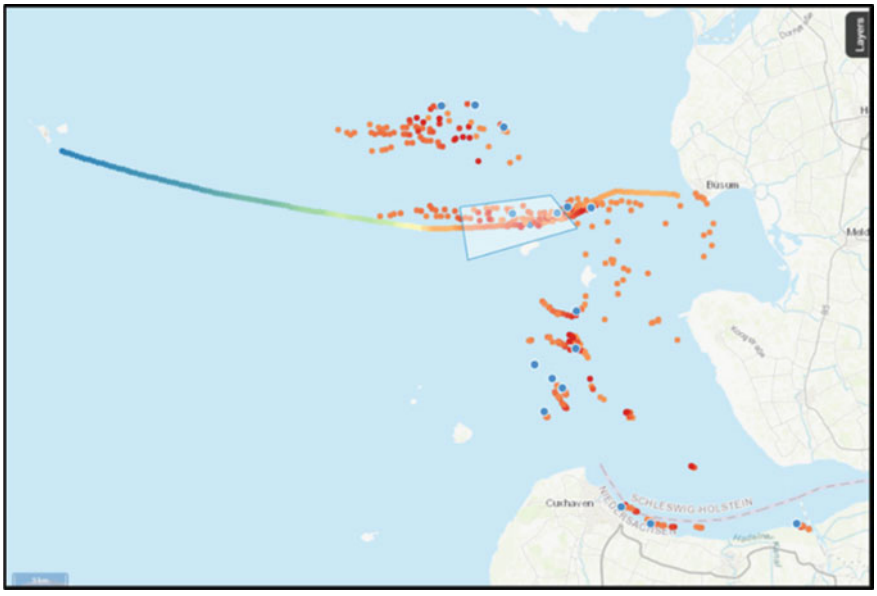

Fig. 4.9 Screenshot of the River Plume Workflow's interactive map. The selected region marks the area of the suspected river plume on the FerryBox transect on 23 June 2013. *Blue dots* highlight one water body's modelled trajectory originating in the Elbe River

origin. The generated time series of chlorophyll gives an overview of productivity changes in the region during and after the flood event. This helped identify promising study regions in the North Sea regarding noteworthy biological events, e.g. unusual algae blooms after a riverine flood event. In general, the River Plume Workflow helped to give better understanding of the sequence of events of the Elbe river flood of 2013 and its impacts on the marine environment of the North Sea.

For more information on the River Plume Workflow, please see https://digitalearth-hgf.de/results/workflows/flood-event-explorer/#accordion-4.

4.6 Conclusions

All examples provided in this chapter have in common that they apply advanced computational approaches to gain new insights from existing data and help to further our process understanding through innovative data analysis. Regardless of the scientific question or exact context, the presented computational approaches address problems that occur frequently in modern Earth and Environmental sciences.

They typically fall into one or more of the following groups:

- There are not enough data available to solve the problem at hand. Modern computational methods can enable scientists to derive information from the combination of related data and provide more context for the scientific question.
- In contrast, some scientific workflows face the opposite problem, namely an abundance of available data. In these cases, insightful data analysis cannot be achieved with classical approaches due to data size and distributed storage. The use of algorithms for automatically classifying events and providing context can greatly improve scientists' process understanding.
- For some scientific questions, the way to a better process understanding involves the comparison or combination of different kinds of existing data. The difficulties here lie in the fact that different datasets of interest are usually not fully comparable in terms of spatial and/or temporal resolution or method of measurement. Computational methods such as the ones described here can take these differences into account and therefore ensure meaningful and correct scientific results.

The applications presented in this chapter demonstrate how computational data exploration and data analytics can help overcome these common problems. Although there is no one-size-fits-all approach to any of the problems we face, the example applications in this chapter show that modern computational approaches can help handling the current paradoxical situation of having access to more data than can be handled by classical methods, while simultaneously needing to overcome a lack of data in other contexts. Moreover, the examples show that those approaches enable us to create additional benefit from the available data and improve our understanding of the complex system Earth.

Acknowledgements The KORA study was initiated and financed by the Helmholtz Zentrum München—German Research Center for Environmental Health, which is funded by the German Federal Ministry of Education and Research (BMBF) and by the State of Bavaria. Furthermore, KORA research was supported within the Munich Center of Health Sciences (MC-Health), Ludwig-Maximilians-Universität, as part of LMUinnovativ. Since 2000, the MI data collection has been co-financed by the German Federal Ministry of Health and Social Security to provide population-based MI morbidity data for the official German Health Report (see www.gbe-bund.de).

References

Ababou R, Bagtzoglou AC, Wood EF (1994) On the condition number of covariance matrices in kriging, estimation, and simulation of random fields. Math Geol 26:99–133. https://doi.org/10.1007/BF02065878

Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, Ghemawat S, Goodfellow I, Harp A, Irving G, Isard M, Jia Y, Jozefowicz R, Kaiser L, Kudlur M, Levenberg J, Mane D, Monga R, Moore S, Murray D, Olah C, Schuster M, Shlens J, Steiner B, Sutskever I, Talwar K, Tucker P, Vanhoucke V, Vasudevan V, Viegas F, Vinyals O, Warden P, Wattenberg M, Wicke M, Yu Y, Zheng X (2016) Tensorflow: large-scale machine learning on heterogeneous distributed systems

Amato F, Guignard F, Robert S et al (2020) A novel framework for spatio-temporal prediction of environmental data using deep learning. Sci Rep 10:22243. https://doi.org/10.1038/s41598-020-79148-7

Appel KW, Pouliot GA, Simon H, Sarwar G, Pye HOT, Napelenok SL, Akhtar F, Roselle SJ (2013) Evaluation of dust and trace metal estimates from the Community Multiscale Air Quality (CMAQ) model version 5.0. Geosci Model Dev 6:883–899. https://doi.org/10.5194/gmd-6-883-2013

Baschek B, Schroeder F, Brix H, Riethmüller R, Badewien TH, Breitbach G, Brügge B, Colijn F, Doerffer R, Eschenbach C, Frie-drich J, Fischer P, Garthe S, Horstmann J, Krasemann H, Metfies K, Merckelbach L, Ohle N, Petersen W, Pröfrock D, Röttgers R, Schlüter M, Schulz J, Schulz-Stellenfleth J, Stanev E, Staneva J, Winter C, Wirtz K, Wollschläger J, Zielinski O, Ziemer F (2017) The coastal observing system for northern and arctic seas (COSYNA). Ocean Sci 13:379–410. https://doi.org/10.5194/os-13-379-2017

Bieser J, Aulinger A, Matthias V, Quante M, Builtjes P (2011) SMOKE for Europe—adaptation, modification and evaluation of a comprehensive emission model for Europe. Geosci Model Dev 4:47–68. https://doi.org/10.5194/gmd-4-47-2011

BLFS: Bayerisches Landesamt für Statistik: GENESIS Datenbank. https://www.statistikdaten.bayern.de/genesis/online/. Last Accessed on 01 September 2021

BLFU: Bayerische Landesamt für Umwelt: Lufthygienische Landesüberwachungssystem Bayern (LÜB). https://www.lfu.bayern.de/luft/immissionsmessungen/messwertarchiv/index.htm. Last Accessed on 01 September 2021

Breitbach G, Krasemann H, Behr D, Beringer S, Lange U, Vo N, Schroeder F (2016) Accessing diverse data comprehensively—CODM, the COSYNA data portal. Ocean Sci 12:909–923. https://doi.org/10.5194/os-12-909-2016

Callies U, Kreus M, Petersen W, Voynova YG (2021) On using Lagrangian drift simulations to aid interpretation of in situ monitoring data. Front Mar Sci 8:769. https://doi.org/10.3389/fmars.2021.666653

Chen K, Breitner S, Wolf K, Hampel R, Meisinger C, Heier M, Von Scheidt W, Kuch B, Peters A, Schneider A (2019) Temporal variations in the triggering of myocardial infarction by air temperature in Augsburg, Germany, 1987–2014. Eur Heart J 40:1600–2160. https://doi.org/10.1093/eurheartj/ehz116

CMEMS North Atlantic Chlorophyll (Copernicus-GlobColour) from Satellite Observations: Daily Interpolated (Reprocessed from 1997). Copernicus Monitoring Environment Marine Service (CMEMS). Available at https://resources.marine.copernicus.eu/product-detail/OCEANCOLOUR_ATL_CHL_L4_REP_OBSERVATIONS_009_098/. Accessed 21 September 2021

CMEMS Atlantic-European North West Shelf-Ocean Physics Reanalysis. Copernicus Monitoring Environment Marine Service (CMEMS). Available at https://resources.marine.copernicus.eu/product-detail/NWSHELF_MULTIYEAR_. Accessed 21 September 2021

Franco B, Mahieu E, Emmons LK, Tzompa-Sosa ZA, Fischer EV, Sudo K, Bovy B, Conway S, Griffin D, Hannigan JW, Strong K, Walker KA (2016) Evaluating ethane and methane emissions associated with the development of oil and natural gas extraction in North America. Environ Res Lett 11:044010. https://doi.org/10.1088/1748-9326/11/4/044010

Greinert J, Schoening T (n.d.) RV POSEIDON Fahrtbericht/Cruise Report POS526—SeASOM: Semi-Autonomous Subsurface Optical Monitoring for methane seepage and cold-water coral studies in the North Sea, Bergen (Norway)—Dogger Bank (Netherlands)—Hirtshals (Denmark)—Tisler (Norway)—[WWW Document]. Report. https://doi.org/10.3289/geomar_rep_ns_51_2019

Holle R, Happich M, Löwel H, Wichmann HE (2005) KORA—A research platform for population based health research. Gesundheitswesen 67:19–25. https://www.doi.org/10.1055/s-2005-858235

Jacob D, Petersen J, Eggert B, Alias A, Bossing Christensen O, Bouwer LM, Braun A, Colette A, Deque M, Georgievski G, Georgopoulou E, Gobiet A, Menut L, Nikulin G, Haensler A, Hempelmann N, Jones C, Keuler K, Kovats S, Kröner N, Kotlarski S, Kriegsmann A, Martin E, Van Meijgaard E, Moseley C, Pfeifer S, Preuschmann S, Rademacher C, Radtke K, Rechid

D, Rounsevell M, Samuelsson P, Somot S, Soussana JF, Teichmann C, Valentini R, Vautard R, Weber B, Yiou P (2014) EURO-CORDEX: new high-resolution climate change projections for European impact research. Reg Environ Change 14(2):563–578. https://doi.org/10.1007/s10113-013-0499-2

Janssens-Maenhout G, Crippa M, Guizzardi D, Muntean M, Schaaf E, Dentener F, Bergamaschi P, Pagliari V, Olivier JGJ, Peters JAHW, van Aardenne JA, Monni S, Doering U, Petrescu AMR, Solazzo E, Oreggioni GD (2019) EDGAR v4.3.2 global atlas of the three major greenhouse gas emissions for the period 1970–2012. Earth Syst Sci Data 11:959–1002. https://doi.org/10.5194/essd-11-959-2019

Janssens-Maenhout G, Pagliari V, Guizzardi D, Muntean M (2012) Global emission inventories in the Emission Database for Global Atmospheric Research (EDGAR) Manual (I) Gridding: EDGAR emissions distribution on global gridmaps. European Commission—Joint Research Centre—Institute for Environment and Sustainability

Marien L, Valizadeh M, Zu Castell W, Nam C, Rechid D, Schneider A, Meisinger C, Linseisen J, Wolf K, Bouwer LM (2022) Machine learning models to predict myocardial infarctions from past climatic and environmental conditions. Nat Haz Earth Syst Sci Disc. https://doi.org/10.5194/nhess-2021-389

NRW. Open Geo Data, State of North-Rhine Westphalia. https://www.opengeodata.nrw.de/produkte/. Last Aaccessed on 01 September 2021

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O et al (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12, 2825–2830

Prill F, Reiner D, Rieger D, Zängl G, Schröter J, Förstner J, Werchner S, Weimer M, Ruhnke R, Vogel B (2019) ICON model tutorial. Working with the ICON model, practical exercises for NWP mode and ICON-ART. Deutscher Wetterdienst, Karlsruhe Institute of Technology, Max-Planck-Institut für Meteorologie

Putter H, Young GA (2001) On the effect of covariance function estimation on the accuracy of kriging predictors. Bernoulli 7(3):421–438. https://projecteuclid.org/euclid.bj/1080004758

Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer, LNCS, Vol. 9351: 234–241. Available at arXiv:1505.04597

Saunois M, Bousquet P, Poulter B, Peregon A, Ciais P, Canadell JG, Dlugokencky EJ, Etiope G, Bastviken D, Houweling S, Janssens-Maenhout G, Tubiello FN, Castaldi S, Jackson RB, Alexe M, Arora VK, Beerling DJ, Bergamaschi P, Blake DR, Brailsford G, Brovkin V, Bruhwiler L, Crevoisier C, Crill P, Covey K, Curry C, Frankenberg C, Gedney N, Höglund-Isaksson L, Ishizawa M, Ito A, Joos F, Kim H-S, Kleinen T, Krummel P, Lamarque J-F, Langenfelds R, Locatelli R, Machida T, Maksyutov S, McDonald KC, Marshall J, Melton JR, Morino I, Naik V, O'Doherty S, Parmentier F-JW, Patra PK, Peng C, Peng S, Peters GP, Pison I, Prigent C, Prinn R, Ramonet M, Riley WJ, Saito M, Santini M, Schroeder R, Simpson IJ, Spahni R, Steele P, Takizawa A, Thornton BF, Tian H, Tohjima Y, Viovy N, Voulgarakis A, van Weele M, van der Werf GR, Weiss R, Wiedinmyer C, Wilton DJ, Wiltshire A, Worthy D, Wunch D, Xu X, Yoshida Y, Zhang B, Zhang Z, Zhu Q (2016) The global methane budget 2000–2012. Earth Syst Sci Data 8:697–751. https://doi.org/10.5194/essd-8-697-2016

Schröter J, Rieger D, Stassen C, Vogel H, Weimer M, Werchner S, Förstner J, Prill F, Reinert D, Zängl G, Giorgetta M, Ruhnke R, Vogel B, Braesicke P (2018) ICON-ART 2.1: a flexible tracer framework and its application for composition studies in numerical weather forecasting and climate simulations. Geosci Model Dev 11:4043–4068. https://doi.org/10.5194/gmd-11-4043-2018

Sieck K, Nam C, Bouwer LM, Rechid D, Jacob D (2020) Weather extremes over Europe under 1.5 and 2.0 °C global warming from HAPPI regional climate ensemble simulations. Earth Syst Dyn 12(2):457–468. https://doi.org/10.5194/esd-2020-4

Singh D, Singh B (2020) Investigating the impact of data normalization on classification performance. Appl Soft Comput 97(Part B):105524. ISSN 1568-4946. https://doi.org/10.1016/j.asoc.2019.105524

Van Dingenen R, Crippa M, Anssens-Maenhout G, Guizzardi D, Dentener F (2018) Global trends of methane emissions and their impacts on ozone concentrations. JRC Sci Policy Rep. https://doi.org/10.2760/8201755

Visschedijk AJH, Denier Van Der Gon HAC, Doornenbal HC, Cremonese L (2018) Methane and ethane emission scenarios for potential shale gas production in Europe. Adv Geosci 45:125–131. https://doi.org/10.5194/adgeo-45-125-2018

Vlasenko A, Matthias V, Callies U (2021) Simulation of chemical transport model estimates by means of neural network using meteorological data. Atmos Environ. https://doi.org/10.1016/j.atmosenv.2021.118236

Volfová A, Šmejkal M (2012) Geostatistical methods in R. Geoinformatics FCE CTU 8:29–54. https://doi.org/10.14311/gi.8.3

Wagenaar D, De Jong J, Bouwer LM (2017) Multi-variable flood damage modelling with limited data using supervised learning approaches. NHESS 17(9):1683–1696. https://doi.org/10.5194/nhess-17-1559-2017

Wessel M, Brandmeier M, Tiede D (2018) Evaluation of different machine learning algorithms for scalable classification of tree types and tree species based on sentinel-2 data. Remote Sens 10:1419. https://doi.org/10.3390/rs10091419

Xie P, Li T, Liu J, Du S, Yang X, Zhang J (2020) Urban flow prediction from spatiotemporal data using machine learning: a survey. Inf Fusion 59: 1–12. ISSN 1566–2535. https://doi.org/10.1016/j.inffus.2020.01.002

Yacovitch TI, Daube C, Herndon SC (2020) Methane emissions from offshore oil and gas platforms in the Gulf of Mexico. Environ Sci Technol 54:3530–3538. https://doi.org/10.1021/acs.est.9b07148