CMUG Extension Phase 2 Deliverable

Reference: D5.1.1.v1: WP5.1 interim progress report
Due date: May 2024
Submission date: May 2024
Version: 1.1.0

Climate Modelling User Group (CMUG)

Extension Phase 2 Deliverable D5.1.1.v1: WP5.1 WP5.1: Machine Learning to advance climate model evaluation and process understanding – interim progress report

Centres providing input: DLR

Version	Date	Status
1.0.0	6 June 2024	first version
1.1.0	21 June 2024	Update with definition of PC algorithm and submit to ESA

Table of Contents

WP5.1: Machine Learning to advance climate model evaluation and process understanding – interim progress report

Summary

This report summarizes the progress made within WP5.1 until May 2024. The main aim of WP5.1 is to develop and apply machine learning (ML) techniques for advanced climate model evaluation and process understanding with ESA CCI data. Progress is reported for the two main tasks "Enhancing observational products for climate model evaluation with machine learning" (WP5.1.1) and "Causal model evaluation for cloud regimes and land cover types" (WP5.1.2) separately.

Within WP5.1.1, a ML-method to derive cloud classes from coarse resolution data such as climate model output has been developed and evaluated with coarse grained ESA Cloud_cci data. Similar to Zantedeschi et al. (2019), cloud type labels from CloudSat are used in a first step to derive the cloud type from collocated physical cloud properties from MODIS using a deep neural network. In a second step, a random forest (RF) regression model is trained on a coarse-grained version of these data to allow for deriving cloud class distributions from coarse-resolution data such as climate model output. This work is documented in the journal article Kaps et al. (2023). The two-stage ML-method from Kaps et al. (2023) has then been applied to 35 years of ESA Cloud_cci data to generate the new *Cloud Class Climatology* dataset (CCClim). As a proof of concept, CCClim is compared to output from a simulation with the ICON-A climate model. This work is summarized in Kaps et al. (accepted).

In WP5.1.2 ESA CCI data are used to better understand and to quantify the main drivers determining observed cloud properties. We apply causal discovery to investigate the links between cloud properties such as cloud cover, cloud water path, cloud top pressure and cloud optical depth and so-called cloud controlling factors, i.e., quantities that impact cloud formation and evolution (e.g., sea surface temperature and amount of available water vapour). For this, causal networks are calculated from time series of daily ESA CCI and ERA5 data. Causal discovery belongs to the field of unsupervised machine learning and aims to discover and quantify causal interdependencies and dynamical links inside a system such as the Earth's climate (Runge et al., 2015; 2019). This approach goes beyond correlation-based measures by systematically excluding common driver effects and indirect links. This is work in progress. This interim report describes the data preprocessing and method used as well as some first results for application to marine stratocumulus clouds.

1 WP5.1.1 – Enhancing observational products for climate model evaluation with machine learning

In WP5.1.1, an approach based on machine learning is developed and applied to derive cloud classes from high-resolution satellite data and coarse-resolution climate models. The aim is to allow for an improved evaluation of clouds in climate models by analysing cloud properties by cloud type. This enables evaluation of the different underlying processes driving formation and evolution of these cloud types in climate models. The method is then applied to ESA Cloud_cci data that are coarse-grained to the resolution of a typical climate model. As a proof of concept, the resulting timeseries of cloud class information from ESA Cloud_cci is then used for comparison with results from a simulation with the ICON-A model (Giorgetta et al., 2018). WP5.1.1 addresses the following two science questions:

- Can cloud classes be derived from ESA Cloud_cci data with machine learning to improve climate model evaluation?
- How well does the ICON-A model reproduce the observed mean properties and variability of satellite derived cloud classes (regime-oriented evaluation)?

1.1 ML-approach to derive cloud class distribution from coarse-resolution data

A two-stage approach based on machine learning has developed to derive cloud classes from high-resolution satellite data and coarse-resolution climate models. This approach is documented in Kaps et al. (2023) in detail and only briefly summarized in the following.

Using cloud type labels from CloudSat and collocated physical cloud properties from MODIS similar to Zantedeschi et al. (2019), cloud type labels can be generated by a deep neural network for cloudy MODIS pixels. The basic approach is shown schematically in *Figure 1*.

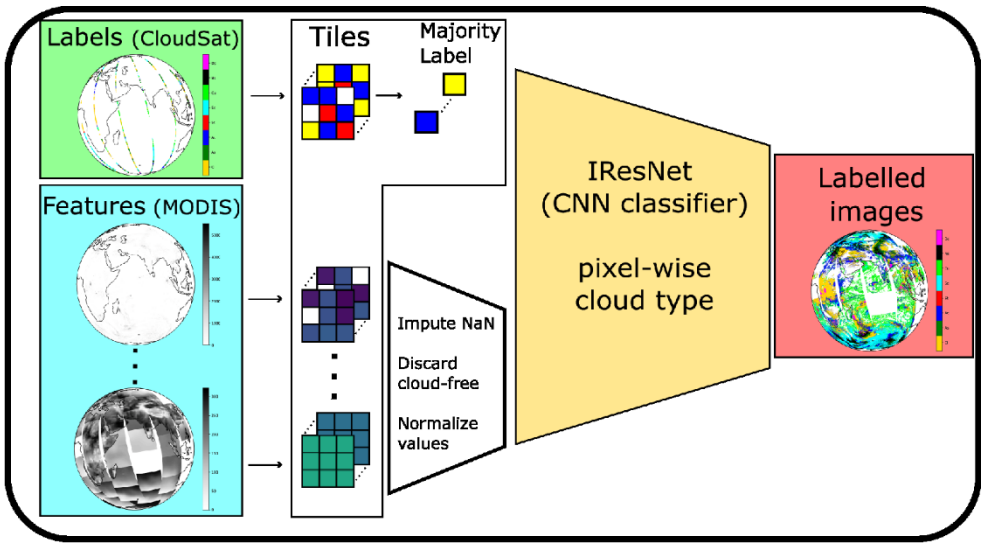

Figure 1: Schematic of the pixelwise classifier, which is a convolutional neural network trained on features from MODIS and one of eight cloud-type labels from CloudSat per pixel. From Kaps et al. (2023).

These data are coarse-grained to the horizontal resolution of a typical climate model of 100 km x 100 km and used to train a random forest (RF) regression model to derive cloud class distributions from coarse-resolution data. This is shown schematically in *Figure 2*.

For details on the method and the datasets used, we refer to Kaps et al. (2023).

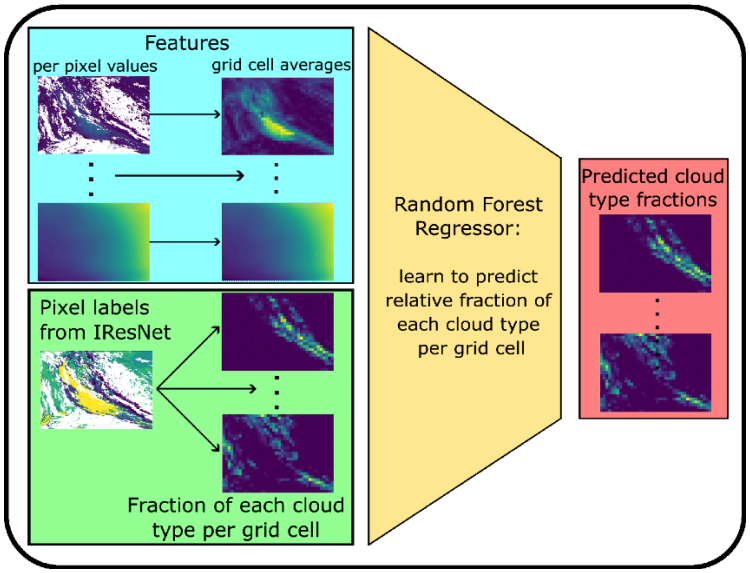

Figure 2: Cloud type predictions for data with low horizontal resolution are obtained by coarse-graining high-resolution predictions as a basis to train a regression model predicting relative amounts of each cloud type for each coarse resolution grid cell. From Kaps et al. (2023).

1.2 Application to ESA Cloud_cci data

The two-stage ML-approach developed in Kaps et al. (2023) is applied to 35 years of ESA Cloud_cci L3U-AVHRR-PM version 3.0 data (Stengel et al., 2020). The dataset contains twice daily measurements from the Advanced Very High Resolution Radiometer (AVHRR) on a 0.05°-grid (L3U data). This is used to generate new "Cloud Class Climatology" dataset (CCClim), which is schematically shown in *Figure 3*.

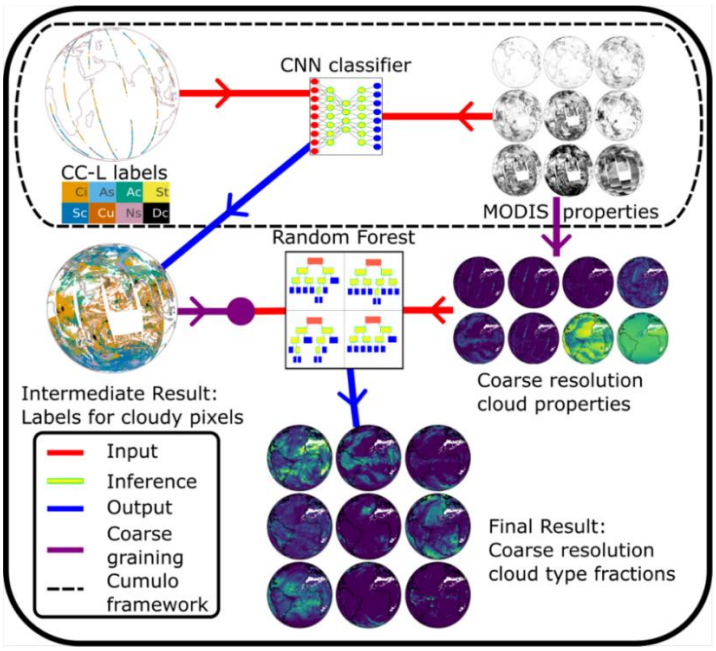

Figure 3: Schematic of the training of the two machine learning models. The second stage is trained on coarse-grained output from the first stage. The trained random forest (RF) is then applied to ESA Cloud_cci *data to generate the CCClim dataset. From Kaps et al. (accepted).*

CCClim contains daily averages of nine cloud-related variables and the relative occurrences of nine classes (eight cloud types + undetermined). CCClim has a global coverage over the time period 1982 through 2016 at a horizontal resolution of 1° x 1° allowing for performing process-oriented analyses of clouds on a climatological time scale. The cloud related variables and cloud types contained in CCClim are listed in *Table 1*. As an example, *Figure 4* shows a time series of the relative frequency of occurrence (RFO) of the eight cloud classes contained in CCClim averaged over the Southern Hemisphere ocean from 1982 through 2016.

A journal article introducing CCClim, more examples of potential scientific applications and a proof of concept comparison to a simulation with the climate model ICON-A has now been accepted for publication in Earth System Science Data (ESSD).

Table 1: List of cloud related variables and cloud types contained in CCClim.

Cloud related variables	Cloud types
cloud water path	Ci: Cirrus/Cirrostratus
ice water path	As: Altostratus
liquid water path	Ac: Altocumulus
cloud optical depth	St: Stratus
effective liquid droplet radius at cloud top	Sc: Stratocumulus
effective ice particle radius at cloud top	Cu: Cumulus
cloud top pressure	Ns: Nimbostratus
surface temperature	Dc: Deep convective
cloud area fraction	

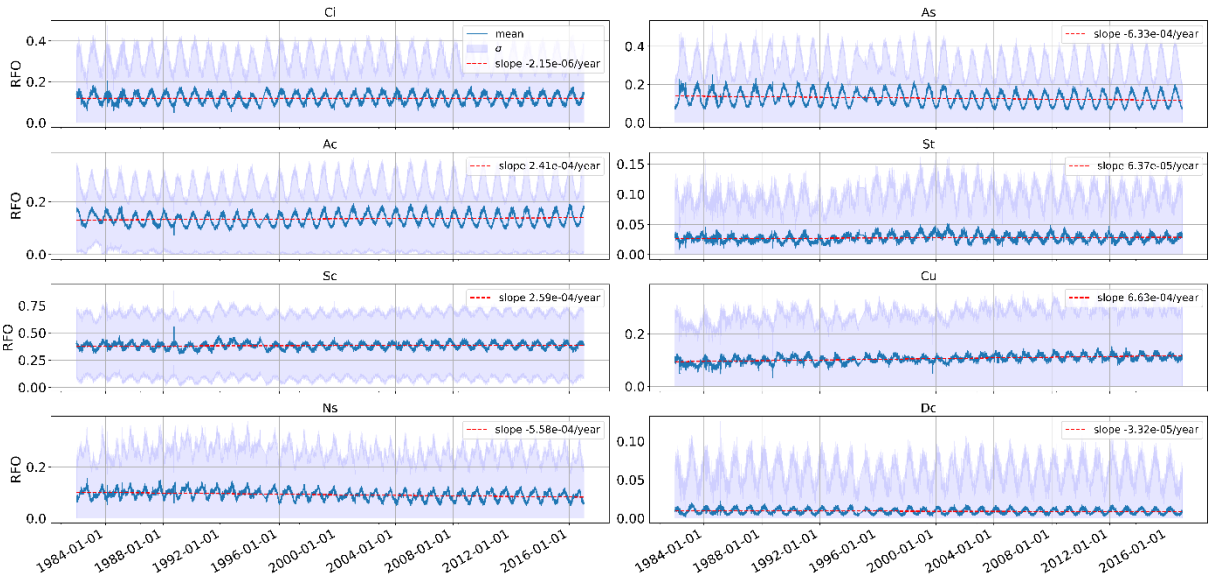

Figure 4: Time series of daily mean relative frequency of occurrence (RFO) with the spatial standard deviation shown as shading for all cloud types averaged over the ocean in the Southern Hemisphere. All types show a consistent seasonal cycle and little anomalies and drift, as shown by the slope of the linear fit. Grid cells with a maximum RFO close to zero (1% quantile) are filtered out. From Kaps et al. (accepted).

2 WP5.1.2 Causal model evaluation for cloud regimes and land cover types

As a key component of the hydrological cycle and the Earth's radiation budget, clouds play an important role in both weather and climate. Our incomplete understanding of clouds and their role in cloud-climate feedbacks leads to large uncertainties in climate simulations. Using causal inference as an unsupervised machine learning method we aim to systematically analyse and quantify causal interdependencies and links between cloud properties and their controlling factors. This approach goes beyond correlation-based measures by systematically excluding common drivers and indirect links. By estimating the causal effect of each of the cloud controlling factors for different cloud regimes we expect to be able to better understand the dominant processes which determine the micro- and macro-physical properties of clouds.

Specifically, causal inference is used to investigate the links between cloud properties and cloud controlling factors, i.e., quantities that impact cloud formation and temporal evolution of the cloud. For this, causal networks are calculated from time series of these variables from satellite and reanalysis datasets averaged over selected geographical regions and cloud regimes in order to quantify the strength of the individual links in the resulting causal graph by applying causal effect estimation.

As a first step we focus on one region, the Pacific Ocean west of South America, where mainly one cloud type is present, marine stratocumulus. The processes controlling marine stratocumulus clouds are already well investigated which allows a better assessment of the ML method. In the following, data and methods are described as well as first results shown. This is work in progress.

2.1 Stratocumulus clouds

Stratocumulus clouds are common over the cooler regions of subtropical and midlatitude oceans where their coverage can exceed 50% in the annual mean (Wood, 2012). 80% of the world's stratocumulus clouds are located over the ocean (Warren et al. 1986, 1988) mainly in eastern subtropical oceans on the western side of the continents and can persist over long time periods (Wood, 2012; Klein and Hartmann, 1993). The low clouds are composed of an ensemble of individual convective elements forming a layer capped by a temperature inversion. The dynamics are primarily driven by convective instability caused by cloud-top radiative cooling (Wood, 2012). A schematic of the most important processes for marine stratocumulus clouds are shown in *Figure 5*.

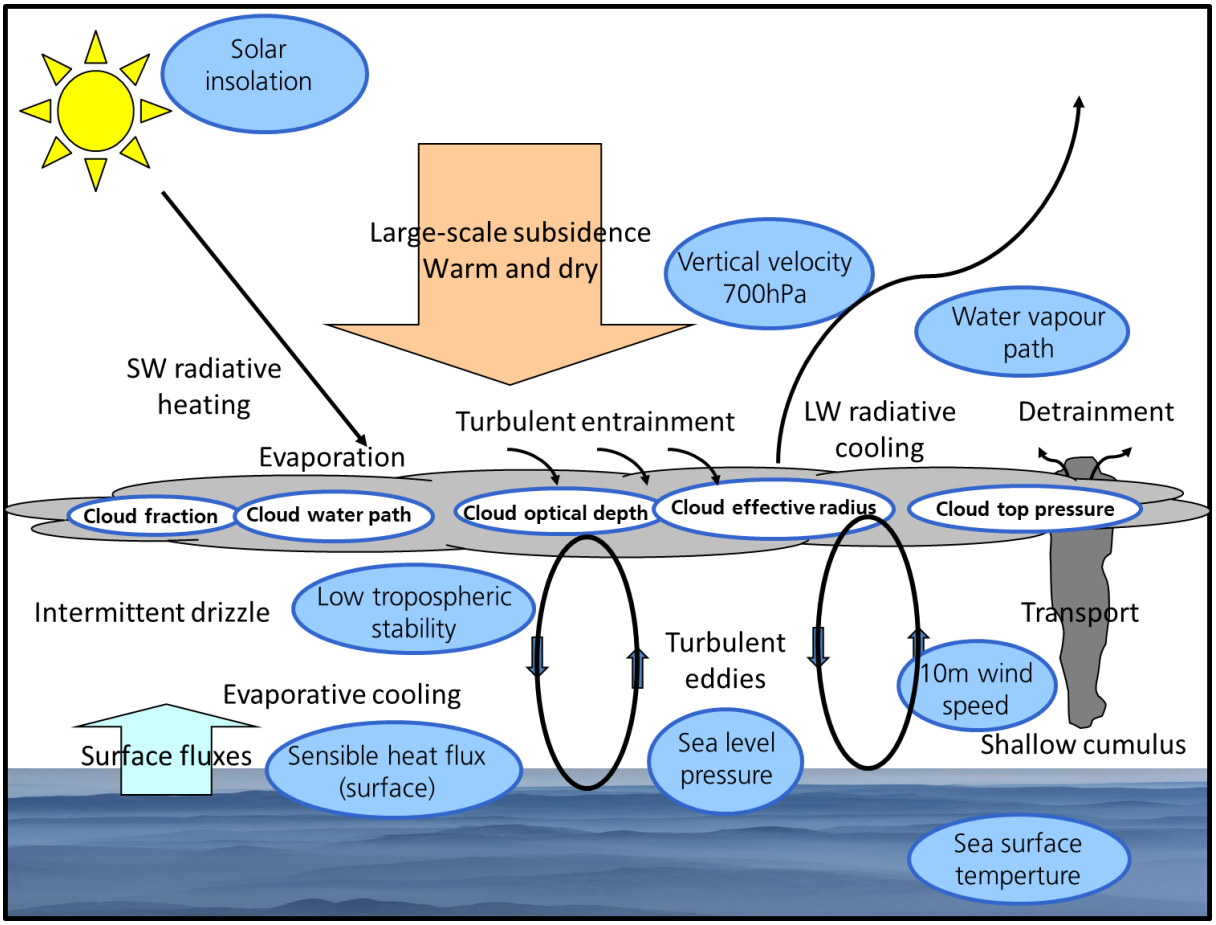

Figure 5: Schematic of important processes for marine stratocumulus. Cloud parameters are shown as white bubbles, cloud controlling factors as blue bubbles.

2.2 Data

We use 5 years (2003-2007) of daily data. The datasets and variables used are summarized in *Table 2*. In cooperation with WP4, reformatting scripts (so-called "CMORizers") were extended or newly written to add support for daily data for these datasets to ESMValTool, which is used to preprocess all data analysed in this study.

Table 2: Datasets and variables used in WP5.1.2.

	Variable	Dataset
Cloud Properties	Total Cloud Fraction (clt), Total Cloud Water Path (clwvi), Cloud Optical Depth (cod), Cloud Effective Radius (reff), Cloud Top Height (ctp)	ESACCI-Cloud (v3.0, L3U, AVHRR-PM, NOAA-16, daily instantaneous data) (Stengel et al., 2020)
Cloud-controlling factors	Sea Surface Temperature Anomaly (tos)	ESACCI-SST (v3.0, Level 4 Analysis product, daily) (Good et al., 2024)
	Water Vapour Path (prw)	ESACCI-Watervapour (CM SAF/CCI TCWV-global (COMBI), v3.1, daily mean data) (Schröder et al., 2023)
	Vertical Velocity at 700hPa (wap700), Lower Tropospheric Stability (LTS), Sea Surface Pressure (psl), Sensible Heat Flux at Surface (hfss), 10m Horizontal Wind Speed (sfcWind)	ERA5 (daily average from hourly data) (C3S, 2017)
	Solar insolation (solin)	CERES-EBAF (monthly mean data) (Loeb et al., 2009)

Figure 6 shows the the marine stratocumulus region over the Southeast Pacific analyzed. We average over 5°× 5° regions because clouds can be assumed to be in equilibrium with their large-scale environment at this horizontal scale (Klein et al.,1995). In order to filter weather-related variability, we apply a low pass filter (Butterworth filter, cut-off= 5 days) (*Figure 7*, blue lines).

In order to increase the data volume as basis for the PCMCI framework (see Section 2.3) we use the timeseries of all 16 5°x5° grid boxes over the Southeast Pacific west of South America (*Figure 6*). The algorithm then calculates the causal links with the information from all 16 regions.

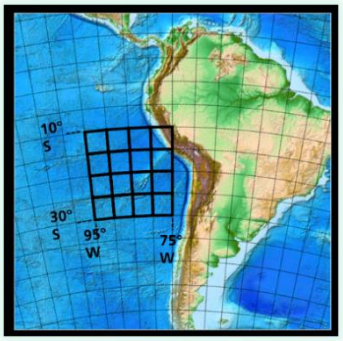

Figure 6: Marine stratocumulus region over the Southeast Pacific analysed in this study consisting of 16 5°x5° boxes (75°- 95°W, 10°- 30°S).

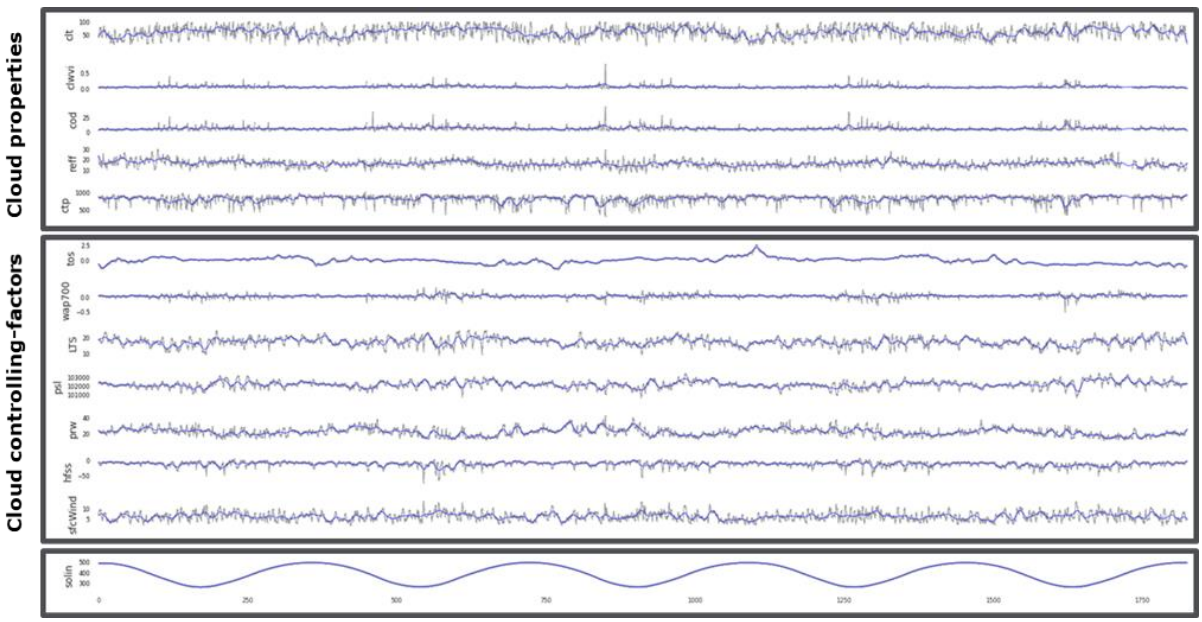

Figure 7: Timeseries of all variables (abbreviations are defined in Table 2) of original daily data (black lines) and after applying a low pass filter to filter out 5-day weather variability (blue lines) averaged over one 5°x5° box over the Southeast Pacific (Figure 6).

2.3 Method

Causal discovery belongs to unsupervised machine learning and aims to discover and quantify causal interdependencies and dynamical links inside a system, such as the Earth's climate (Runge et al., 2015; 2019). This approach goes beyond correlation-based measures by systematically excluding common driver effects and indirect links. For our study we are using the software package TIGRAMITE which is a time series analysis Python module. It is available on GitHub (https://github.com/jakobrunge/tigramite) and it allows to reconstruct graphical models (conditional independence graphs) from discrete or continuously-valued time series based on the PCMCI framework with different methods and conditional independence tests to be chosen. PCMCI consists of two stages: (i) PC_1 condition selection (the Peter-Clark algorithm, named after the original authors) to identify relevant conditions for all timeseries variables and (ii) the momentary conditional independence (MCI) test, where conditional independence between the variables given their estimated parents from the PC step is tested for a chosen significance level (Runge et al., 2019).

As the method we chose PCMCI+ which can identify the full, lagged and contemporaneous, causal graph (up to the Markov equivalence class for contemporaneous links) under the standard assumptions of Causal Sufficiency, Faithfulness and the Markov condition (Runge et al., 2020). The "+" in PCMCI+ means that beside lagged dependencies also contemporaneous links are considered. In a first attempt we apply robust partial correlation (RobustParCorr) as conditional independence test which is valid for linear dependencies, including non-Gaussian distributions. It transforms the data to a normal distribution prior to the partial correlation test. After including solar insolation, it turns out that the dependencies are not linear anymore which is the reason for switching to the conditional independence test CMIknn (Runge et al., 2018). This is a conditional mutual information test based on nearest-neighbour estimator. No assumptions about the parametric form of the dependencies are required as these can be directly estimated from the underlying joint density.

TIGRAMITE also has a "Causal Effect" class that allows to estimate (conditional) causal effects and mediation based on assuming a causal graph. The aim is to use this approach to quantify the causal strength of the cloud controlling factors on the cloud properties.

2.4 First results

As a first step of the PCMCI framework in TIGRAMITE we investigate the distributions of all variables and the dependencies between the variables. As an example, *Figure 8* shows the kernel density estimates, the joint densities between all variables and the marginal distribution of each variable on the diagonal. The distributions are non-Gaussian but mainly reasonable linear. Using initially the RobustParCorr as conditional independence test, we also explore CMIknn as an option (see Section 2.3 for more details).

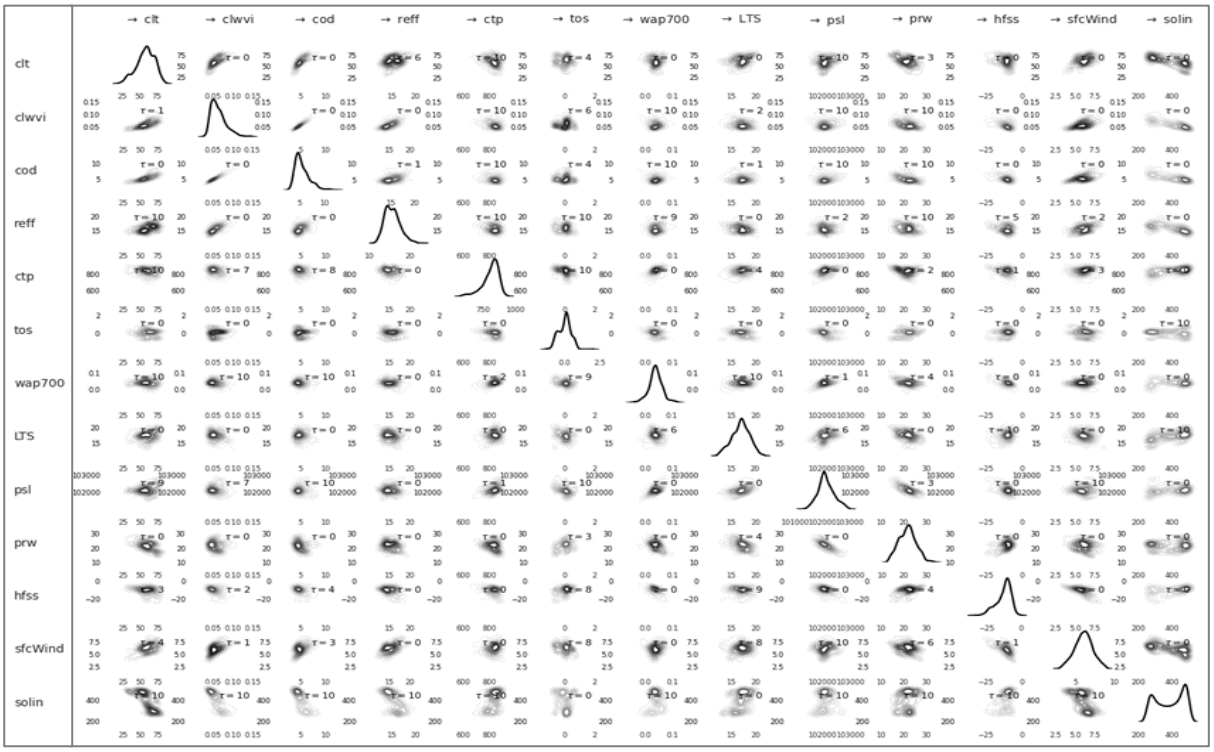

Figure 8: Kernel density estimates of the joint and marginal (diagonal panels) densities for one 5°x5° grid box (see Figure 6).

For applying PCMCI+ together with the conditional independence test RobustParCorr on the data, we use a maximum time lag of 1 day and the significance level for all tests pc_alpha is 0.1. The resulting causal graph is shown in *Figure 9*. Conflicting links, i.e. links with conflicting directions coming out of the orientation phase of existing links, are marked with crosses at the end mean. Application of the Causal Effect class to quantify the strength of the links in the causal graph is still work in progress.

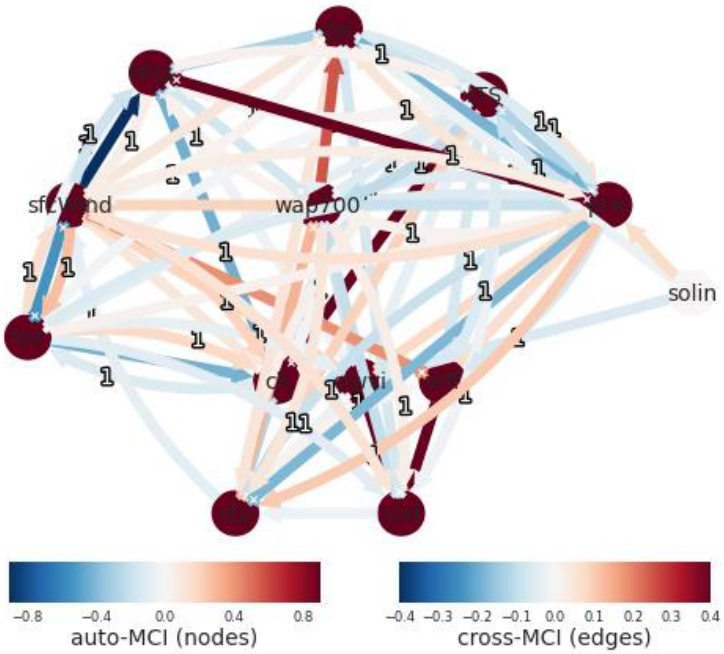

Figure 9: Process graph of causal links. The link colours refer to the cross-MCI value and the node colour denotes the auto-MCI value at the lag with maximum absolute value.

2.5 Further steps

The aim of this work is to get a fully directed causal graph. The idea is to move to CMIknn as conditional independence test instead of RobustParCorr. This might solve some of the current problems in the calculations. In a next step, we will add known connections as link assumptions before calculating the causal links. Setting known links and their direction or removing links found by the algorithm that are known to have no physical basis are expected to help the algorithm. The aim is to obtain a fully directed causal graph with which we can estimate direct and mediated causal effects in the causal graph and thus quantify the influence of cloud controlling factors on observed cloud properties.

The ultimate goal is then to apply this method also to other selected cloud regimes, e.g. clustered by specific land cover types.

References

C3S (2017): ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate, https://cds.climate.copernicus.eu/cdsapp#!/home.

Giorgetta, M. A., Brokopf, R., Crueger, T., Esch, M., Fiedler, S., Helmert, J., et al. (2018). ICON-A, the atmosphere component of the ICON Earth system model: I. Model description. Journal of Advances in Modeling Earth Systems, 10, 1613–1637. doi: 10.1029/2017MS001242.

Good, S.A.; Embury, O. (2024): ESA Sea Surface Temperature Climate Change Initiative (SST_cci): Level 4 Analysis product, version 3.0. NERC EDS Centre for Environmental Data Analysis, 09 April 2024. doi: 10.5285/4a9654136a7148e39b7feb56f8bb02d2.

Kaps, A., Lauer, A., Camps-Valls, G., Gentine, P., Gómez-Chova, L., & Eyring, V. (2023). Machine-Learned Cloud Classes From Satellite Data for Process-Oriented Climate Model Evaluation. Ieee Transactions on Geoscience and Remote Sensing, 61, https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10025016.

Kaps, A., Lauer, A., Kazeroni, R., Stengel, M., & Eyring, V. (accepted). Characterizing clouds with the CCClim dataset, a machine learning cloud class climatology. Earth System Science Data.

Klein, S. A. et al. (1995): On the relationships among low-cloud structure, sea surface temperature, and atmospheric circulation in the summertime Northeast Pacific. J. Clim., 8, 1140–1155.

Loeb, N. G. et al. (2009): Toward optimal closure of the Earth's top-of-atmosphere radiation budget. J. Climate, 22, 748-766, doi: 10.1175/2008JCLI2637.1.

Runge, J., Petoukhov, V., Donges, J. et al. (2015): Identifying causal gateways and mediators in complex spatio-temporal systems. Nat Commun 6, 8502, doi: 10.1038/ncomms9502.

Runge, J. (2018): Conditional Independence Testing Based on a Nearest-Neighbor Estimator of Conditional Mutual Information. In Proceedings of the 21st International Conference on Artificial Intelligence and Statistics. http://proceedings.mlr.press/v84/runge18a.html.

Runge, J. et al. (2019): Detecting and quantifying causal associations in large nonlinear time series datasets.Sci. Adv.5, eaau4996, doi: 10.1126/sciadv.aau4996.

Runge, J. (2020): Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence, UAI 2020,Toronto, Canada, 2019, AUAI Press, 2020. http://auai.org/uai2020/proceedings/579_main_paper.pdf

Schröder, M. et al. (2023): A combined high resolution global TCWV product from microwave and near infrared imagers - COMBI, Satellite Application Facility on Climate Monitoring, doi: 10.5676/EUM_SAF_CM/COMBI/V001, https://doi.org/10.5676/EUM_SAF_CM/COMBI/V001.

Stengel, M., Stapelberg, S., Sus, O., Finkensieper, S., Wurzler, B., Philipp, D., et al. (2020). Cloud_cci Advanced Very High Resolution Radiometer post meridiem (AVHRR-PM) dataset version 3: 35-year climatology of global cloud and radiation properties. Earth System Science Data, 12(1), 41-60. <Go to ISI>://WOS:000505955500001

Stengel, M. et al. (2020): Cloud_cci Advanced Very High Resolution Radiometer post meridiem (AVHRR-PM) dataset version 3: 35-year climatology of global cloud and radiation properties, Earth Syst Sci Data Earth Syst Sci Data, 12, 41–60.

Wood, R. (2012): Stratocumulus Clouds. Mon. Wea. Rev., 140, 2373–2423, doi: 10.1175/MWR-D-11-00121.1.

Zantedeschi, V., Falasca, F., Douglas, A., Strange, R., Kusner, M. J., & Watson-Parris, D. (2019). Cumulo: A Dataset for Learning Cloud Classes. Paper presented at the NeurIPS 2019 Workshop Tackling Climate Change with Machine Learning.