Climate Modelling User Group [CMUG]

Deliverable D2.0h

Interim Progress Report on WP5.8: Using Machine Learning to Evaluate and Understand our Capability to Model Tropical Wetland Methane Emissions

Centres providing input: University of Leicester, Met Office.

Version nr.	Date	Status
0.1	20 Nov. 2024	Input from partners
1.0	29 Nov. 2024	First complete version

Deliverable D2.0h

Interim Progress Report on WP5.8: Using Machine Learning to Evaluate and Understand our Capability to Model Tropical Wetland Methane Emissions

Contents

Interim Progress Report on WP5.8: Using Machine Learning to Evaluate and Understand our Capability to Model Tropical Wetland Methane Emissions

1. Purpose and scope of this report

This document summarises the progress on WP5.8 ("Using Machine Learning to Evaluate and Understand our Capability to Model Tropical Wetland Methane Emissions") of the CCI CMUG project. The study aims to enhance our understanding of tropical wetland methane emissions and derive useful insights to help us improve the models. This activity has a strong technical element, with the use of machine-learning emulators of land surface model JULES combined with CCI satellite-based datasets in an innovative model-data fusion approach. The emulators are thus used to generate a new dataset, which is evaluated against atmospheric inversions of GHG-CCI data, and to leverage the benefits of explainable AI to explore how the input data drive the resulting model output.

There are three partners involved in this activity: the University of Leicester lead the project and are in charge of emulator development and testing; the Met Office are responsible for the JULES simulations; and the University of Edinburgh are responsible for the atmospheric methane inversions.

We report here the progress on the following tasks:

- Production of ensemble of JULES wetland methane simulations.
- Initial steps in emulator development.
- Considerations for the use of CCI datasets with the emulator, including dataset selection, gap filling and best practices.

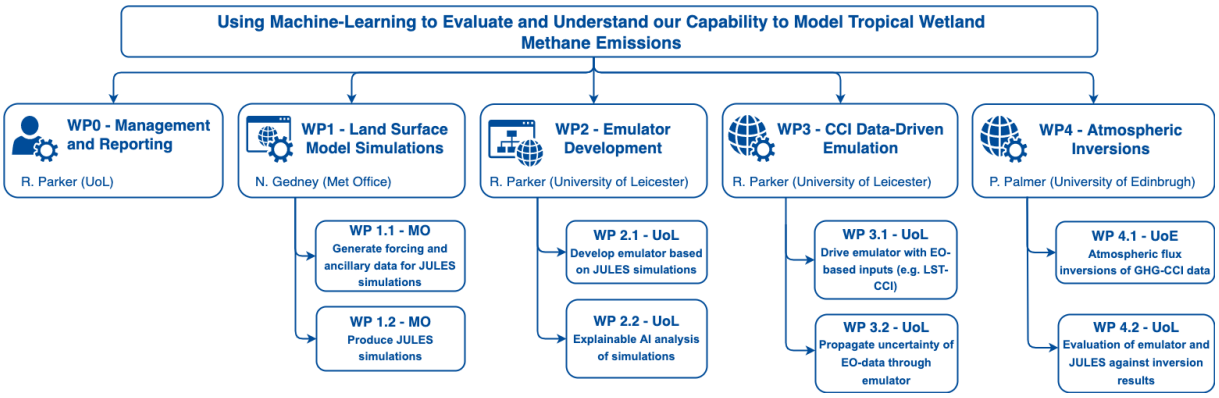

Figure 1. Project structure diagram.

2. WP5.8.1 Land Surface Model Simulations

The aim of this work package is to generate an ensemble of JULES wetland methane simulations focused over tropical Africa, covering a wide range of configurations and input data, that we can use for training and testing the emulator.

2.1 The JULES model

JULES (Joint UK Land Environment Simulator; Clark et al, 2011; Best et al, 2011; Gauci et al 2021) is a land surface model which models the carbon, energy and water cycles. It has a simple groundwater model that simulates a grid box mean water table depth, which, when combined with the grid box statistical distribution of topographic index, simulates the distribution of water table within the grid box and thus the fraction of inter-fluvial inundation, fw.

JULES simulates the methane emitted (F_{CH4}) from the inundated fraction of the grid box. The default version of JULES used applies one of the following equations depending on whether soil carbon (Cs) or root exudates (which are assumed to be proportional to net primary productivity (NPP)) are assumed to be the available substrate for methanogenesis:

$$F_{CH4}(Cs) = fw.Kcs.Cs.Q10(Cs,T)^{\left(\frac{(T-To)}{10}\right)}$$
$$F_{CH4}(NPP) = fw.Knpp.NPP.Q10(NPP,T)^{\left(\frac{(T-To)}{10}\right)}$$

where T (K) is the mean top 1 m soil temperature and To a reference temperature (273.16 K), K_{cs} and K_{npp} are global constants tuned to produce an appropriate global total wetland flux and the Q10 factors describe the amounts by which reaction rates increase with a 10 K temperature increase. Here $Q10(NPP,T) = Q10(NPP)^{\left(\frac{To}{T}\right)}$.

All the simulations carried out so far have calculated the methane emissions based on root exudates/NPP.

2.2 Forcing and ancillary data (WP5.8.1.1)

JULES is forced with observation-based meteorology, and requires sub-daily near-surface air temperature, wind speed and humidity, and short-wave and long-wave radiation, and precipitation. It also requires land ancillary data for soil properties and topographic index (Marthews et al., 2015), as well as LAI and canopy height (as we are fixing vegetation cover in these simulations) and spatial maps of the vegetation fractions for the different plant functional types.

Two different historical observational datasets are considered: CRU/CRU-JRA55 (Harris, 2014) and GSWP3-W5E5 (Dirmeyer et al., 2006; Kim 2017; Lange, 2019; Cucchi et 849 al., 2020).

The fractions of the JULES plant functional types (PFTs) are produced from the aggregation of land cover types into surface tile types from International Geosphere-Biosphere Programme (IGBP) maps (Section 2.1). LAI and canopy height are prescribed based for each plant functional type.

Standard JULES uses sand-silt-clay percentages from the Harmonized World Soil Database (FAO 2012) to describe the hydraulic soil properties for most soils (Best et al 2011). However, these formulations do not represent heavily leached, tropical, kaolinitic soils well (Tomasella & Hodnet, 1997). This is due to the micro-aggregation of their soil particles resulting in a hybrid behaviour, with the properties of both sand and clay. Some of our JULES simulations use soil ancillaries which have been generated to include these tropical soils. The Harmonized World Soil Database is used to determine in which grid boxes these soils (oxisols and/or ultisols – USDA soil taxonomy) are dominant. In these grid boxes the formulae developed in Tomasella & Hodnet (1998) are applied to generate the soil properties.

2.3 JULES simulations (WP5.8.1.2)

Table 1 shows the ensemble of JULES simulations carried out, and Figures Figure 2 and Figure 3 show example maps of average wetland fraction and methane emissions, respectively, for June 2016, generated for all ensemble members. Simulation 1 is taken as the control. Simulation 2 is carried out to investigate how JULES responds to different driving data and uncertainty in the observational driving data.

Temperature is a key driver of wetland methane emissions but there is significant uncertainty in Q10 (Gedney et al., 2019). Simulations 3 and 4 investigate the temperature sensitivity to methane production and its likely lower and upper bounds (Gedney et al., 2019). Incorporation of oxisols and ultisols dramatically impacts the water table depth and therefore the inundation extent, which is another key factor controlling emissions. Simulations 5 & 6 investigate the importance of including these different soil hydraulic properties.

Priorities for the next set of simulations could include looking at the impact of using soil carbon as the methane substrate rather than root exudates, different parameterisation in the soil hydrology module and potentially more complex wetland methane emission models.

Table 1. Ensemble of JULES simulations produced for this study.

Sim. No.	Sim. ID	Forcing data	$Q10$	Soil properties
1	u-dc921	GSWP3	$Q10(NPP) = 1.6$	Standard soils: sand-silt-clay
2	u-dc910	CRUJRA	$Q10(NPP) = 1.6$	Standard soils: sand-silt-clay
3	u-de834	GSWP3	$Q10(NPP) = 2.3$	Standard soils: sand-silt-clay
4	u-de835	GSWP3	$Q10(NPP) = 1.3$	Standard soils: sand-silt-clay
5	u-ck917	GSWP3	$Q10(NPP) = 1.6$	Includes oxisols
6	u-ck843	GSWP3	$Q10(NPP) = 1.6$	Includes oxisols and ultisols

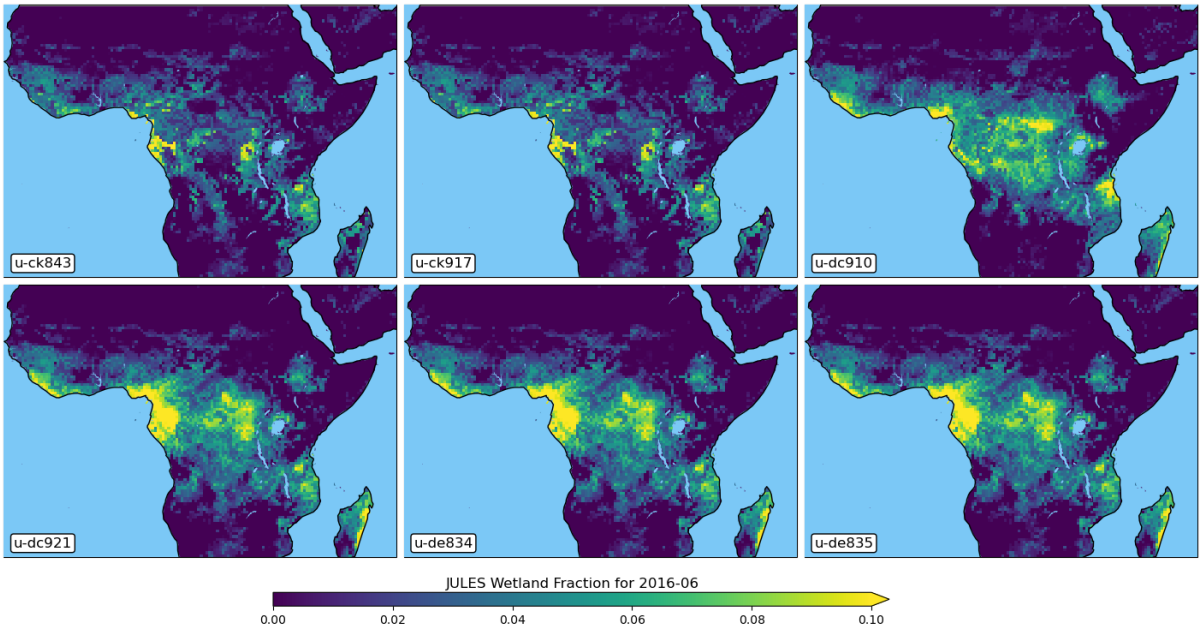

Figure 2. Example wetland fraction generated in the JULES ensemble of simulations.

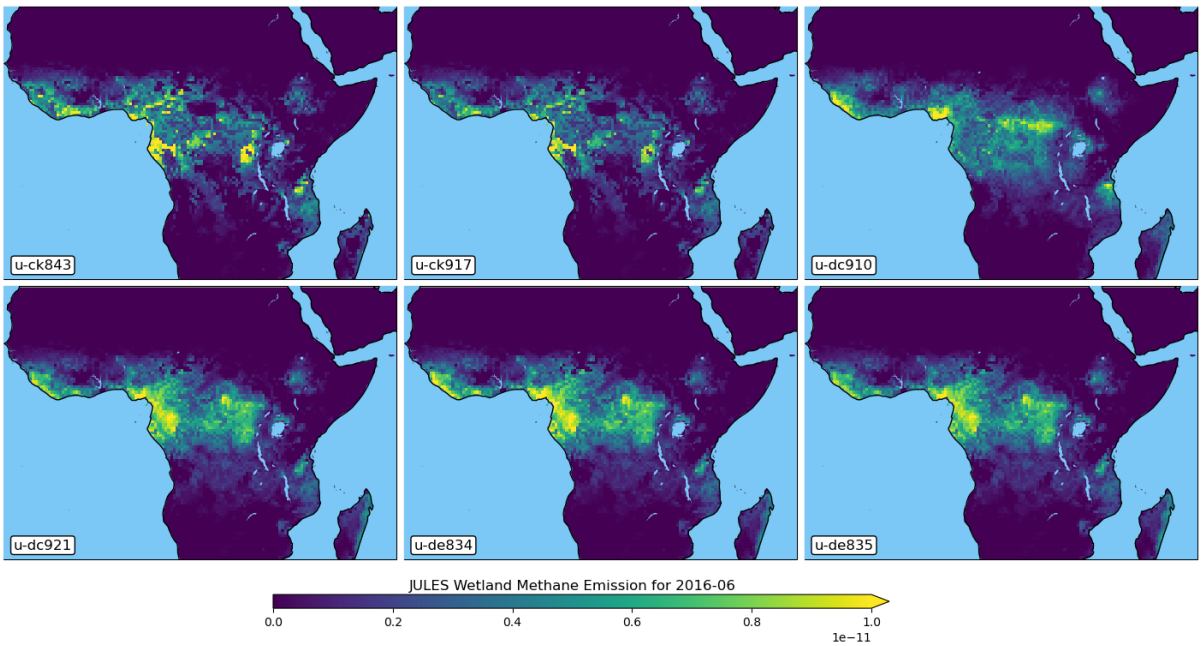

Figure 3. Example wetland methane emissions generated in the JULES ensemble of simulations.

3. WP5.8.2 Emulator Development

The aim of this work package is to develop a machine-learning emulator of wetland methane from JULES and to leverage its explainable AI capabilities to better understand the model behaviour. This work builds on previous GPP and soil moisture emulators that we developed successfully.

3.1 Python framework to streamline emulator development

Previous work on JULES emulators involved many manual steps, from the data pre-processing to the feature selection, hyperparameter tuning and visualisation, which makes development slow, cumbersome and harder to keep track of iterations. To automate and streamline emulator development, we are building a Python framework, which is modular and model agnostic, and will have capabilities such as automated feature selection to find the most optimal set of input features.

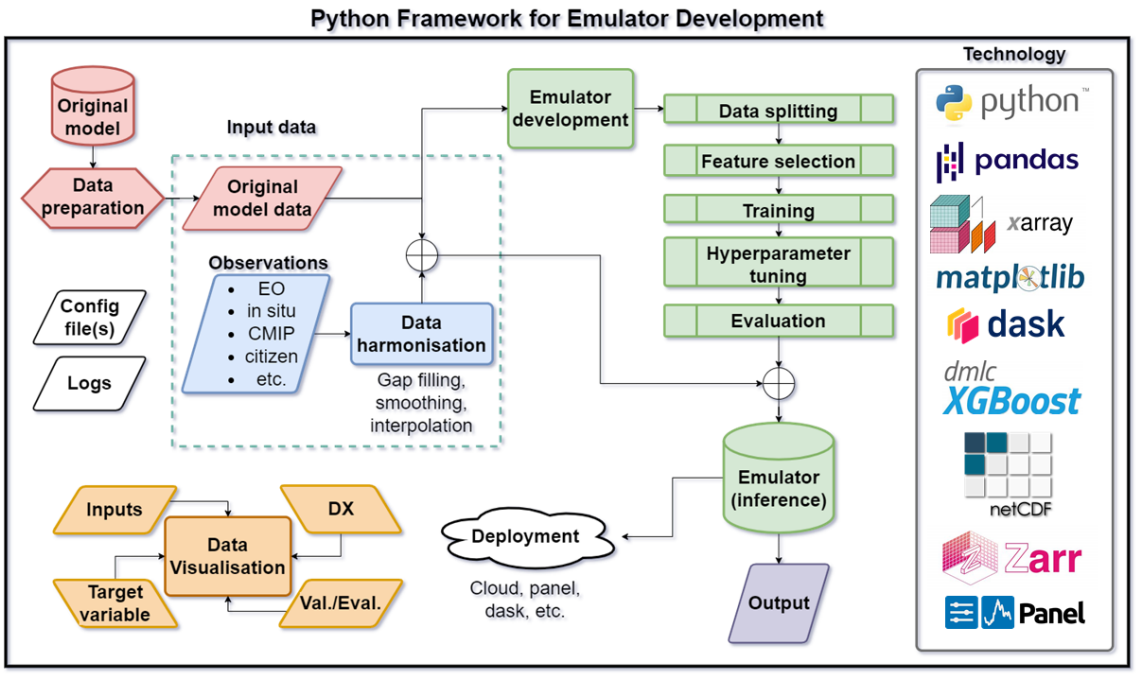

Figure 4. Flow diagram of the emulator development framework.

3.2 Wetland methane emulator development

Emulator development is currently ongoing. We have completed some preparatory work to pre-process the data from the JULES simulations, including converting them into the right format for the emulators and applying scaling factors. The plan is to create two emulators as a two-stage approach: one for wetland extent, and one for wetland methane. Both parameters are intrinsically related and of interest in their own right to answer the wetland science questions.

The emulators are built using a decision-tree based algorithm known as *gradient boosting*, for which we use the XGBoost implementation in Python (Chen and Guestrin, 2016). This algorithm is very popular in data science owing to its flexibility and high performance in a wide range of applications. It works by combining multiple weak models to create a stronger model. In this study, the selection of hyperparameters is based on the tuning we performed as part of our work on previous emulators, and might need slight adjustments for optimal performance in this particular application.

An optimal feature selection is critical for good emulator performance. We start by selecting all the available features that might be relevant for the target variable (e.g. wetland methane), based on our knowledge of the physical processes involved. The initial feature set can be optimised at a later stage to remove inputs that provide redundant information, for example, those which are highly correlated to other inputs. Though largely a 'trial-and-error' exercise, the feature selection optimisation can be partly guided by a feature importance analysis, which provides a summary of which inputs were more relevant in determining the output.

We have started building an emulator of wetland extent and are currently working on the input feature selection. Some of the inputs we have tried so far include: precipitation, surface pressure, specific humidity at 1.5m height, topographic index, total soil moisture in column, etc. However, because the emulator is point-based (each simulated data point is independent of the rest in time and space) it does not have any 'memory', which means that we are missing some important information such as the recent precipitation history. To work around this limitation, we are engineering additional features derived from other inputs, such as monthly means or lagged parameters.

3.3 Next steps

We will continue working on the feature selection for the wetland extent emulator, adding some derived inputs such as multiple lagged precipitation variables. We expect this task to be lengthy and to go through several iterations.

The next step is to start work on the wetland methane emulator, following a similar process to the wetland extent one, including the feature selection.

Once the emulators are finalised, we will also explore explainable AI techniques like SHAP (SHapley Additive exPlanations; Lundberg and Lee, 2017) to learn what influence the different inputs have on the output. This analysis will help us better understand the factors affecting particular predictions of wetland extent or methane emissions, and may also lead to improvements in JULES.

4. WP5.8.3 CCI Data-Driven Emulation

The aim of this work package is to generate a new dataset of wetland methane emissions by driving the emulator developed in WP5.8.2 with ESA-CCI datasets (e.g. land surface temperature), along with an associated characterisation of the uncertainty.

4.1 Data pre-processing preparation

Model vs observational variables

To be able to use observations as inputs, they must be equivalent to the model variables the emulator was trained on. For example, JULES surface temperature (*tstar*) can be considered equivalent to the land surface temperature (LST) observed by satellites, so an emulator trained on *tstar* can be used with LST data. The observational variables also must have a similar distribution to their JULES counterparts to ensure the emulator training covers the whole range of possible values. Therefore, careful consideration needs to be taken for all input features during the feature selection process to ensure there is a suitable observational equivalent, taking into account the different assumptions that might have been made in the model compared to the observational datasets.

Resolution and grids

Different EO datasets will typically have different spatial and temporal resolutions compared to each other (and to JULES), and be sampled on different grids. However, all datasets used for training and running the emulator must have the same spatial and temporal resolutions and grids, so harmonisation of all datasets is needed. Our approach is to select a resolution (e.g. daily, 5 km) and upsample or downsample datasets as required, typically using nearest neighbours for upsampling, and averaging for downsampling.

In theory, the emulator can be trained on data at any spatiotemporal resolution, provided the processes of interest are well represented at that resolution. Then the emulator can be used with EO inputs at a different resolution to the JULES data used for training. This relative independence from resolution allows us to train the emulator using coarser (and thus less computationally intensive) JULES simulations, and benefit from the higher resolution of the EO data.

Gap filling

Remote sensing data can sometimes be rather sparse owing to factors such as cloud cover or quality filtering. However, the emulator requires all input variables to be able to make predictions, which means that, even if there is a single input missing, it will not generate an output. Therefore, to maximise the number of predictions, it is necessary to carry out some gap filling. In previous work with a similar emulator, we performed some basic gap filling by doing temporal linear interpolation, with a maximum gap of 30 data points. Figure 5 shows an example of one day of ESA CCI soil moisture data over Europe where observations are very

sparse (Figure 5a), and the resulting data after gap filling (Figure 5b), with much improved coverage.

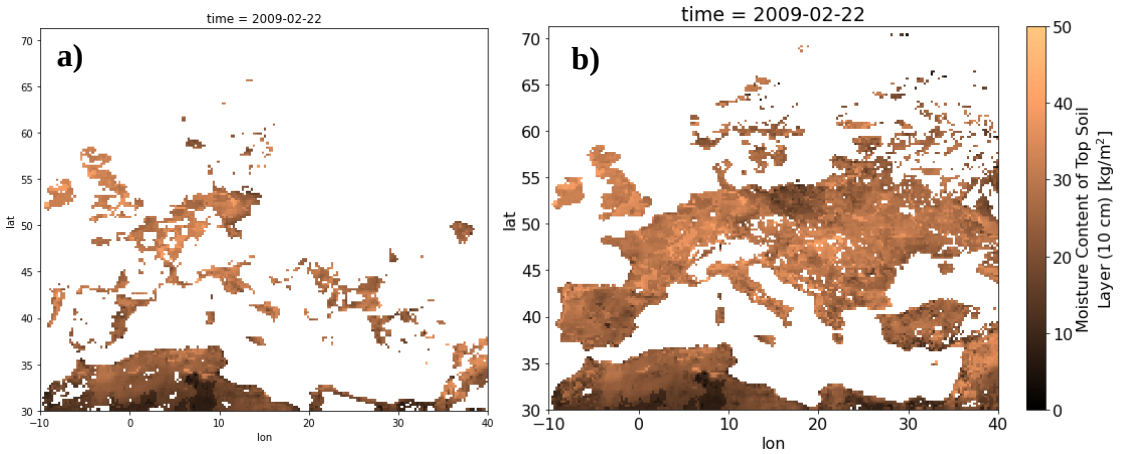

Figure 5. Example of gap filling for CCI soil moisture dataset for 22/02/2009: a) Original data. b) Gap-filled data. Some gaps remain where there are more than 30 missing data points between observations, but coverage is much improved.

Land cover classes to PFT conversion: crosswalk table

One of the EO datasets that we will be using in this study is the ESA CCI Land Cover product. However, the land cover classes used in the CCI product (the Land Cover Classification System, LCCS) are not directly equivalent to JULES' PFTs, so we need to define a crosswalk table to convert between them. When creating such a table, assumptions need to be made which will inevitably amplify uncertainties. For example, JULES has six different types of grasses, including the differentiation between C3 and C4, which are not present in LCCS. In a crosswalk table we created for a previous emulator (Figure 6, adapted from Poulter et al., 2015), we divided the LCCS grass fractions equally among the six JULES grass PFTs. Other assumptions include the exclusion of certain types of vegetation not present in the study domain (e.g. tropical trees in Europe).

Since we created the crosswalk table in Figure 6, there have been new efforts to map LCCS to PFTs (e.g. Wang et al., 2023) that might help refine our table for this study. In addition, a new CCI land cover product that uses PFTs instead of LCCS classes will become available soon, as explained by Céline Lamarche at the Mid-Term Review meeting for this activity. This new product can potentially be very useful for this study and we will include it in the list of datasets to consider in this work package when the emulator is finalised.

C3S LCCS dataset		JULES PFTs												
lc_type	lc_name	0	1	2	3	4	5	6	7	8	9	10	11	12
		Broadleaf Deciduous	Broadleaf Evergreen Tropical	Broadleaf Evergreen Temperate	Needleleaf Deciduous	Needleleaf Evergreen	C3 grasses	C3 grasses crops	C3 grasses pastures	C4 grasses	C4 grasses crop	C4 grasses pastures	Shrubs Deciduous	Shrubs Evergreen
0	no_data													
10	cropland_rainfed						16.67	16.67	16.67	16.67	16.67	16.67		
11	cropland_rainfed_herbaceous_cover						16.67	16.67	16.67	16.67	16.67	16.67		
12	cropland_rainfed_tree_or_shrub_cover						8.33	8.33	8.33	8.33	8.33	8.33	50	
20	cropland_irrigated						16.67	16.67	16.67	16.67	16.67	16.67		
30	mosaic_cropland	5		5			12.50	12.50	12.50	12.50	12.50	12.50	5	10
40	mosaic_natural_vegetation	5		5			10.83	10.83	10.83	10.83	10.83	10.83	10	15
50	tree_broadleaved_evergreen_closed_to_open			90									5	5
60	tree_broadleaved_deciduous_closed_to_open	70					2.50	2.50	2.50	2.50	2.50	2.50	15	
61	tree_broadleaved_deciduous_closed	70					2.50	2.50	2.50	2.50	2.50	2.50	15	
62	tree_broadleaved_deciduous_open	30					5.83	5.83	5.83	5.83	5.83	5.83	25	
70	tree_needleleaved_evergreen_closed_to_open					70	2.50	2.50	2.50	2.50	2.50	2.50	5	10
71	tree_needleleaved_evergreen_closed					70	2.50	2.50	2.50	2.50	2.50	2.50	5	10
72	tree_needleleaved_evergreen_open					30	5.00	5.00	5.00	5.00	5.00	5.00	5	5
80	tree_needleleaved_deciduous_closed_to_open				70		2.50	2.50	2.50	2.50	2.50	2.50	5	10
81	tree_needleleaved_deciduous_closed				70		2.50	2.50	2.50	2.50	2.50	2.50	5	10

Figure 6. Snippet of the crosswalk table used with previous emulators, which was adapted from Poulter et al. (2015).

Discussion with CCI product leads

EO datasets are complex, involve many assumptions, and multiple versions might be available. To ensure we use the most suitable EO products, and that they are used correctly, we engaged with the science leads for key CCI products that we anticipate will be needed in this study. Leads for soil moisture, land cover, greenhouse gases (GHG), and land surface temperature (LST) attended our Mid-Term Review meeting on 11th of November 2024 and provided updates on their products, as well as helpful information and advice for our project.

Michael Buchwitz, from CCI-GHG, recommended using the CH4_S5P_WFMD (v1.8) product, and clarified that potential biases from low-albedo wetland areas and spurious methane signals are being addressed and there would be further improvements in early 2025.

Darren Ghent, from CCI-LST, explained that the best products are those from a single sensor, especially MODIS and SLSTR, and they should be used when possible. He also mentioned that his team do sophisticated gap filling on the LST data as an intermediary step for downscaling, taking into account land cover, and even though this gap-filled product is not publicly available, they can share it with us when needed.

Wouter Dorigo, from CCI-Soil Moisture, mentioned that there is a new version of the product, covering 45 years of data. He explained that they produce a dataset version that has been gap filled using a sophisticated machine-learning technique, though most of the filling taking place in northern regions, where data quality is poorer.

Céline Lamarche, from CCI-Land Cover, highlighted that there is a new version of the product that uses PFTs instead of LCCS classes, which is a much better solution to the LCCS-PFT mapping than the crosswalk table we have used in the past (Figure 6).

The discussion was very productive and CCI science leads provided useful recommendations and advice that will be instrumental in this work package. We learned about new products and developments that will improve essential steps in our data processing, such as gap filling and LCCS-to-PFT mapping. We will continue to engage with the data providers throughout the project to ensure we apply best practices when using their data.

4.2 Next steps

Although we have started some preparatory work to understand the various EO datasets available and how to use them, the main part of this work package requires the output of WP5.8.2, that is, the trained emulators. Once we generate them, the next steps will be to:

- Finalise the list of EO datasets required.
- Continue the conversation with the data producers.
- Define common spatial and temporal resolutions to harmonise the datasets.
- Perform data pre-processing and harmonisation.
- Drive emulators with pre-processed EO datasets.
- Propagate uncertainty from EO datasets through emulator to estimate overall uncertainty of predictions for wetland extent and methane emissions.
- Perform flux inversions of the CCI-GHG methane data and evaluate JULES and EO-driven emulator against them (WP5.8.4).

5. Summary

The aim of this activity is to gain a better understanding of drivers and responses of tropical wetland methane emissions by using a machine-learning emulator of the JULES land surface model trained across a wide ensemble of simulations and leveraging explainable AI capabilities. The activity involves colleagues from University of Leicester, University of Edinburgh, and the Met Office.

Progress so far includes the generation of six JULES ensemble members covering a range of scenarios with different driving data and wetland-related parameter settings; the development of a Python framework for automated emulator training and testing; initial work on feature selection and engineering for a wetland extent emulator; and preliminary work on EO data selection and usage considerations, including a very fruitful discussion with ESA CCI science leads for LST, GHG, soil moisture and land cover.

Next steps include the generation of further JULES ensemble members with more complex scenarios and emissions models; further work on emulator feature selection, engineering and training; analysis of emulator outputs using explainable AI; generation of new datasets using the emulator and the EO data; and evaluation of our data against flux inversions of CCI-GHG.

6. References

Best MJ, Pryor M, Clark DB, Rooney GG, Essery RLH, Ménard CB, et al. The joint UK land environment simulator (JULES), model description–Part 1: energy and water fluxes. Geosci. Model Dev. 2011 4, 677–699. doi:10.5194/gmd-4-677-2011.

Chen, T., & Guestrin, C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, 785–794. https://doi.org/10.1145/2939672.2939785.

Clark DB, Mercado LM, Sitch S, Jones CD, Gedney N, Best MJ et al. The joint UK land environment simulator (JULES), model description– Part 2: carbon fluxes and vegetation dynamics. Geoscientific Model Dev. 2011 doi: 10.5194/gmd-4-701-2011.

Cucchi, M., Weedon, G. P., Amici, A., Bellouin, N., Lange, S., Müller Schmied, H.,Hersbach, H. and Buontempo, C.: 2542 WFDE5: bias-adjusted ERA5 reanalysis data for impact studies. Earth System Science Data, 12, 2097–2120, 2020.

Dirmeyer, P. A., Gao, X., Zhao, M., Guo, Z., Oki, T., and Hanasaki, N. GSWP-2: Multimodel analysis and impli- cations 2595 for our perception of the land surface. Bulletin of the American Meteorological Society, 87(10):1381–1398, 2006.

Gauci V, Figueiredo V, Gedney N, Pangala SR, Stauffer T, Weedon GP et al. Non-flooded riparian Amazon trees are a regionally significant methane source. Phil. Trans. of Royal Soc. A. 2021 doi: 10.1098/rsta.2020.0446.

FAO/IIASA/ISRIC/ISSCAS/JRC, 2012. Harmonized World Soil Database (version 1.2). FAO, Rome, Italy and IIASA, Laxenburg, Austria.

Harris, I., Jones, P.D., Osborn, T.J. and Lister, D.H.: Updated high-resolution grids of monthly climatic observations – the 2735 CRU TS3.10 Dataset, Int. J. Climatol., 34: 623-642. https://doi.org/10.1002/joc.3711, 2014.

Kim H., Global Soil Wetness Project Phase 3 Atmospheric Boundary Conditions (Experiment 1) [Data set]. Data 2896 Integration and Analysis System (DIAS)., https://doi.org/10.20783/DIAS.501, 2017.

Lange S., WFDE5 over land merged with ERA5 over the ocean (W5E5). V. 1.0. 2019. doi:10.5880/pik.2019.023, 2019.

Lundberg, Scott M., and Lee, Su-In. "A Unified Approach to Interpreting Model Predictions." Advances in Neural Information Processing Systems. Curran Associates, Inc., 2017.

Marthews TR, Dadson SJ, Lehner B, Abele S, Gedney N. High-resolution global topographic index values for use in large-scale hydrological modelling. Hydrol. Earth Syst. Sci. 2015 19:91–104.

Poulter, B., MacBean, N., Hartley, A., Khlystova, I., Arino, O., Betts, R., Bontemps, S., Boettcher, M., Brockmann, C., Defourny, P., Hagemann, S., Herold, M., Kirches, G., Lamarche, C., Lederer, D., Ottlé, C., Peters, M., and Peylin, P.: Plant functional type classification for earth system models: results from the European Space Agency's Land Cover Climate Change Initiative, Geosci. Model Dev., 8, 2315–2328, https://doi.org/10.5194/gmd-8-2315-2015, 2015.

Tomasella, J & Hodnet, MG. Estimating unsaturated hydraulic conductivity of Brazilian Soils using soil-water retention data. Soil Science. 1997 162(10): 703-712.

Tomasella, J & Hodnet, MG. Estimating soil water retention characteristics from limited data in Brazilian Amazonia. *Soil Science.* 1998 163(3): 190-202.

Wang, L., Arora, V. K., Bartlett, P., Chan, E., and Curasi, S. R.: Mapping of ESA's Climate Change Initiative land cover data to plant functional types for use in the CLASSIC land model, Biogeosciences, 20, 2265–2282, https://doi.org/10.5194/bg-20-2265-2023, 2023.

7. Glossary

Acronyms	
AI	Artificial Intelligence
CCI	Climate Change Initiative
CMUG	Climate Modelling Users Group
GHG	Greenhouse Gases
GPP	Gross Primary Productivity
JULES	Joint UK Land Environment Simulator
LCCS	Land Cover Classification System
LST	Land Surface Temperature
ML	Machine Learning
NPP	Net Primary Productivity
PFT	Plant Functional Type
SHAP	SHapley Additive exPlanations