,
Horizon 2020 Societal challenge 5:
Climate action, environment, resource
efficiency and raw materials

VERIFY

Observation-based system for monitoring and verification of greenhouse gases

GA number 776810, RIA

Deliverable number (relative in WP)	**D3.2**
Deliverable name:	Second state of the art database
WP / WP number:	3
Delivery due date:	Month 26 (March 2020)
Actual date of submission:	Month 27 (April 2020)
Dissemination level:	Public
Lead beneficiary:	University of Aberdeen
Responsible	Matthias Kuhnert
Contributor(s):	Matthias Kuhnert, Pete Smith, Matthew McGrath, Philippe Peylin, Karina Winkler, Richard Fuchs, Martin Herold, Emanuele Lugato, Richard Engelen, Adrian Leip, Philippe Ciais, Mart-Jan Schelhaas, Gert-Jan Nabuurs, Frank Deneter, Ronny Lauerwald, Pierre Regnier, Are Olsen, Maximilian Reuter, Meike Becker
Internal reviewer:	Philippe Peylin, Aurélie Paquirissamy

Changes with respect to the DoA
None
Dissemination and uptake (Who will/could use this deliverable, within the project or outside the project?)
The collected data are freely available (in some case registration is/will be necessary). In the project these data build the basis for model simulations in WP3 and WP4. The web-pages for data download are listed in section 3. The provided data build the basis for running the models as well as for calibration and validation of the bottom up models.
Short Summary of results (<250 words)
The state of the art database addresses the data demand by the different modeling groups in order to provide data for forcing, calibration and evaluation of the models. We were already able to summarize most relevant datasets in deliverable D3.1 (the first version of the database deliverable). Therefore, this new version only resumes previous datasets and describes in details the new ones (since D3.1 was submitted). New datasets concern coastal ocean fluxes and biomass and to a certain level land use/land cover data and XCO2. For these datasets only the methodology was provided in the initial deliverable. Some of the other datasets got also replaced by better or more refined data (i.e., climate data). The available data of coastal ocean carbon fluxes cover all European shelf areas in a 6 hour time step for the period 1998-2018. Given the need for the VERIFY modeling groups, a monthly resolution was picked for analysis. The land use/land cover data are provided over an extended period (now 1900-2015). Additionally changes in the approach are reported. The climate data are provided in a finer resolution, which also affects the calculation of the management timing for croplands. All details are reported in the core of the deliverable. All data are up-loaded on the VERIFY THREDDS server (accessible through http://verify.lsce.ipsl.fr/index.php/products). Section 3 of the deliverable resume again the access to all datasets.
Evidence of accomplishment (report, manuscript, web-link, other)
All the datasets will be accessible though the VERIFY web site and the dedicated data-products page: http://verify.lsce.ipsl.fr/index.php/products Note that some of these data are password protected during a consolidation phase and thus only accessible to the VERIFY partners (accessible through the internal share-point platform). Most data have been uploaded on the portal and are accessible through a catalogue and a THREDDS server (see section 3).

Version	Date	Description	Author (Organisation)
V0	21/03/2020	Creation/Writing	Matthias Kuhnert (University of Aberdeen)
V0.1	21/03/2020	Writing/Formatting/Delivery	Matthias Kuhnert, Pete Smith, Matthew McGrath, Philippe Peylin, Karina Winkler, Richard Fuchs, Martin Herold, Emanuele Lugato, Adrian Leip, Philippe Ciais, Mart-Jan Schelhaas, Gert-Jan Nabuurs, Frank Deneter, Ronny Lauerwald, Pierre Regnier, Are Olsen, Maximilian Reuter, Heike Becker
V1	08/04/2020	Formatting/Delivery on the Participant Portal	Matthias Kuhnert, Pete Smith, Matthew McGrath, Philippe Peylin, Aurelie Paquirissamy

1. Glossary

Abbreviation / Acronym	Description/meaning
AWB	agricultural waste burning
COSCAT	COastal Segmentation and the related CATchments
CRU	Climatic Research Unit
EEA	European Environment Agency
EMEP	The co-operative programme for monitoring and evaluation of the long-range transmission of air pollutants in Europe (inofficially 'European Monitoring and Evaluation Programme' = EMEP)
ESA	European Space Agency
ESDAC	European Soil Data Center
FAO	Food and Agriculture Organization
HILDA	Historic Land Dynamics Assessment
HWSD	Harmonized World Soil Database
LULC	Land use/Land Cover
NOAA	National Oceanic and Atmospheric Administration
RMSE	Root mean square error
UERRA	Uncertainties in Ensembles of Regional ReAnalysis

2. Executive Summary

This deliverable provides an update to the state-of-the-art database available in VERIFY. As outlined in task T3.1, this database serves all models across WP3 and WP4, providing input and verification data for both process-based and data-driven models focused on the land surface and the atmosphere. Along with deliverable D3.1 (the first version of the database deliverable), this deliverable provides the most complete and up-to-date description the VERIFY database, describing improvements and new data sets.

Despite most datasets being already complete, the project's collection, aggregation, development and modification of data is still ongoing. The main parts of this deliverable cover data related to climate, land use/land cover, terrestrial biomass and coastal CO_2 emissions. Additionally, there are modifications to crop management data and atmospheric mole fraction of CO_2 (XCO2). The land use/land cover data set evolves constantly throughout the course of the VERIFY project to enable considerations of various project partners. In this deliverable the temporal resolution, temporal coverage, and data map projections have been updated. The climate dataset is being extended to cover the most recent project year. Terrestrial biomass data are new, and while they have not yet been added to the database, they are described here in anticipation. The data for the coastal CO_2 emissions were added to the database recently; the dataset itself and its creation are described here. These two data sets are the only new additions to the database, with the rest fully described in deliverable D3.1.

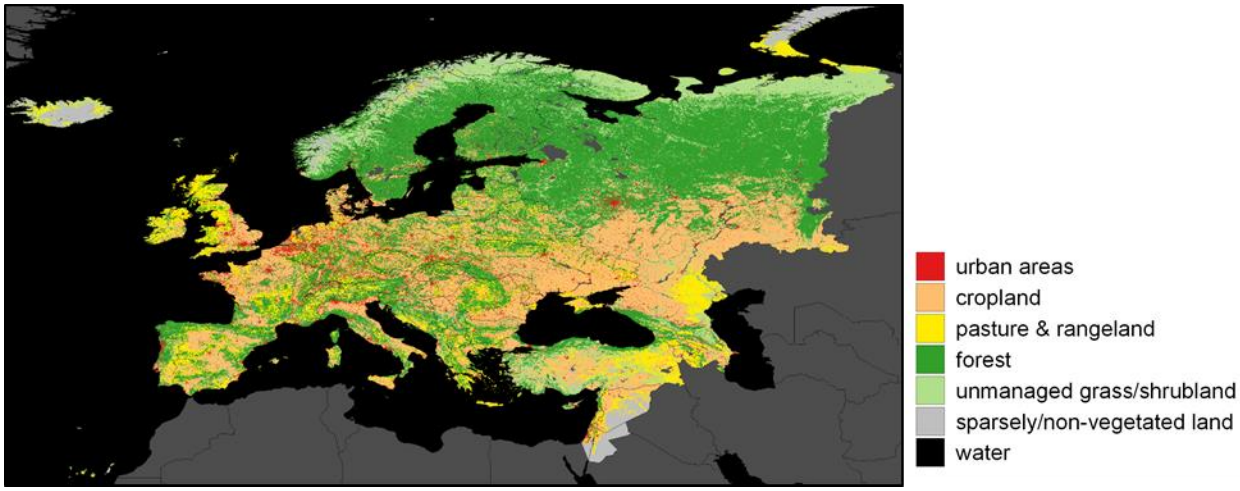

Figure 1 : Land use/cover map (quick view) for Europe in 2015 from HILDA+

Despite the progress made on the land use/land cover and biomass data, the work on these data sets will continue. The actual data are already sufficient for use in VERIFY, but further developments will improve the data. As the data collection for these products is complete (all sections are covered and all data requirements are addressed) further addition of data will only aim at improving data quality and providing additional flexibility for users.

Land use/land cover data in the VERIFY project are now available for the period 1900-2015 with an annual resolution. This is an exceptional data set and, following consultations with project partners, the providers modified it for a longer temporal coverage with a higher temporal resolution. The climate data has been extended to cover the year 2019 to enable VERIFY to maintain operational status. The new datasets of terrestrial biomass and emissions from the European shelf seas address known gaps in the database. With these additions the database is complete and covers all relevant areas for greenhouse gas emission inventories.

3. Introduction

In this deliverable we report progress and changes since the last deliverable D3.1. Table 1 shows the complete list of datasets and contact person to get an overview about the available data. Detailed descriptions are listed on the VERIFY SharePoint platform (for VERIFY partners only). All data are provided to all project partners and will be made available to everyone (when the data is freely available) through a THREDD server (see section 3). The datasets will mainly contribute to the simulation approaches in WP3 and WP4.

Dataset	Name/ model	Inst.	Coverage	Resolution	Time frame	Contact in the project
Land use	CORINE[1]	EEA	global	0.1 km	2000, 2006, 2012	
Soil	HWSD[2]	FAO	global	1 km		
Land use	HILDA[3]	KIT/WU	global	1 km	1900-2015	Richard Fuchs[a], Karina Winkler[b] Martin Herold[c]
Biomass	BIOMASS-CCI[3]	WU	global	0.1 deg	2000, 2010, 2017	Martin Herold[c]
N-Deposition	EMEP model[4]	JRC		1 km	2010	Frank Dentener[d]
Erosion	RUSLE[5]	ESDAC	Europe	100 m	2015	Emanuele Lugato[e]
Soil data	LUCAS[6]	ESDAC	Europe	500 m/ 1 km	2009-2015	Emanuele Lugato[e]
Climate	C3S-ERA-5[7]	ECMWF	global	31 km	2008-current	Richard Engelen[f] Matt McGrath[p]
Climate	UERRA[8]	ECMWF	Europe	10 km	1961-2018	Richard Engelen[f] Matt McGrath[p]
Flux data	FLUXNET[9] network		global	sites	diverse	Dario Papale, Werner Kutsch
Fertiliser application rates	CAPRI	JRC	Europe	0.25°	2000-2012	Adrain Leip[g]
Management timing (crop)		UNIABDN	EU28	0.25°	2000-2015	Matthias Kuhnert[h]

Fresh water fluxes		ULB	Europe	0.1°	2016	Ronny Lauerwald[i]
Ocean coastal fluxes	SOCAT[11]	UiB	Northern Europe	0.125°	1998-2018	Are Olsen[j], Meike Becker[o]
Forest management		WU	Europe	0.125°	2000-2015	M.-J. Schelhaas[k], G.-J. Nabuurs[l]
Grassland management		CEA-LSCE	Europe	0.5°	1860-2012	Philippe Ciais[m]
atmospheric CO2	FOCAL[10]	UBremen	global	2 km	2015-2016	Maximilian Reuter[n]

Table 1 : Available dataset, from the internet (CORINE and HWSD) and provided by project participants.

Email addresses:

a richard.fuchs@kit.edu
b karina.winkler@kit.edu
c martin.herold@wur.nl
d frank.dentener@ec.europa.eu
e Emanuele.LUGATO@ec.europa.eu
f richard.engelen@ecmwf.int
g Adrian.LEIP@ec.europa.eu
h Matthias.kuhnert@abdn.ac.uk
i rlauerwa@ulb.ac.be
j are.olsen@uib.no
k martjan.schelhaas@wur.nl
l gert-jan.nabuurs@wur.nl
m philippe.ciais@lsce.ipsl.fr
n mreuter@iup.physik.uni-bremen.de
o meike.becker@uib.no
p matthew.mcgrath@lsce.ipsl.fr

The data for climate, soil, erosion, freshwater fluxes land use/land cover and atmospheric CO2 are already uploaded on the THREDDS server (WP3-input dataset).

1 https://www.eea.europa.eu/publications/COR0-landcover
2 http://www.fao.org/soils-portal/soil-survey/soil-maps-and-databases/harmonized-world-soil-database-v12/en/
3 https://www.wur.nl/en/Research-Results/Chair-groups/Environmental-Sciences/Laboratory-of-Geo-information-Science-and-Remote-Sensing/Models/Hilda/HILDA-data-downloads.htm
4 http://webdab.emep.int/Unified_Model_Results/
5 https://esdac.jrc.ec.europa.eu/content/soil-erosion-water-rusle2015
6 https://esdac.jrc.ec.europa.eu/resource-type/european-soil-database-soil-properties

7 https://www.ecmwf.int/en/about/media-centre/science-blog/2017/era5-new-reanalysis-weather-and-climate-data
8 https://confluence.ecmwf.int//display/UER
9 https://fluxnet.fluxdata.org/login/?redirect_to=/data/download-data/
10 http://www.iup.uni-bremen.de/~mreuter/TN_XCO2-OCO2-FOCAL_v08.pdf
11 www.socat.info

For some sections there is no progress or changes to show or report. This might be because the final product was already provided for the first deliverable or the ongoing work does not yet show significant changes to the last deliverable. Missing data or rather very preliminary results could be replaced by improved datasets, as was the case for the coastal ocean fluxes. The work on the land use/land cover data and the biomass data is still ongoing, as there will be new datasets developed for VERIFY during the remaining 2 years of the project. The provided climate data are not new compared to the high-resolution products produced for D3.1, but there was a need to extend them for the year 2019 to allow the model to run up to the current year minus one.
Based on the demand and needs of the different groups, all required data are provided so far. Further datasets, which will be provided later or in an improved version, will enhance the model simulation and evaluation, but do not block any progress for ongoing work.

4. Description of the different input and forcing datasets for WP3 activities

4.1. State of the art climate data

In 2019, the VERIFY project, through the combined efforts of the University of East Anglia, ECMWF, and the LSCE, successfully processed high-resolution meteorological forcing data from the UERRA project at three-hour resolution across Europe, including using the observational means from the CRU database to re-align and extend the UERRA dataset back to the year 1901. The choice of the UERRA dataset was made as it satisfied the primary objectives of the VERIFY project: spatial resolution around 10 km, subdaily resolution, temporal coverage to as close to the beginning of the observed 20th-century warming as possible, and an operational status that guaranteed that data for the year 2018 would be available by April 2019. The primary downside of the dataset was the completion of the UERRA project in August 2019, after which production of the dataset would cease.

The current efforts since late fall 2019 have focused on finding a replacement dataset which means all of the original VERIFY requirements. Possible candidates are the CERRA project (the follow-up project to UERRA); EDA; ERA5; WFDE5; and ERA5-Land. Out of all these, only ERA5-Land appears to satisfy the requirements and will be made available in time for the next cycle of VERIFY simulations set to begin during spring of 2020. One challenge is the timing of the VERIFY work and the size of the ERA5-Land files. As ERA-Land is approximately the same spatial resolution as UERRA-HARMONIE but covering a much greater spatial extent and possessing a finer temporal resolution (1h instead of 3h), the files are several times larger, delaying practical applications.

To compensate for this, we plan to begin with the year 2019 for the ERA5-Land data, and work backwards in time. Tests will be done on the ERA5-Land data after re-alignment with the CRU observational dataset to see what discrepancies may occur between one year (January-August 2019) of the UERRA-Harmonie data and one year (January-December 2019) of ERA5-Land, in order to determine the best way to smooth the transition and not cause unrealistic jumps in the model results. As new years of ERA5-Land (2018, 2017, 2016, etc.) are added, the existing year from UERRA-Harmonie will be replaced. In addition, three-hour temporal resolution will be used for 2019 to allow easier merging with UERRA-HARMONIE for the project year 2020, after which we will explore increasing the temporal resolution to one-hour.

4.2. Land use datasets and high resolution land cover change and biomass mapping

4.2.1. High resolution land cover

We provide an updated version (vEUR-0.1) of the European subset of the HILDA+, a global dataset on land use/land cover (LULC) change. We developed the HILDA+ land use/land cover maps using a data-driven reconstruction approach as described in deliverable D3.1.

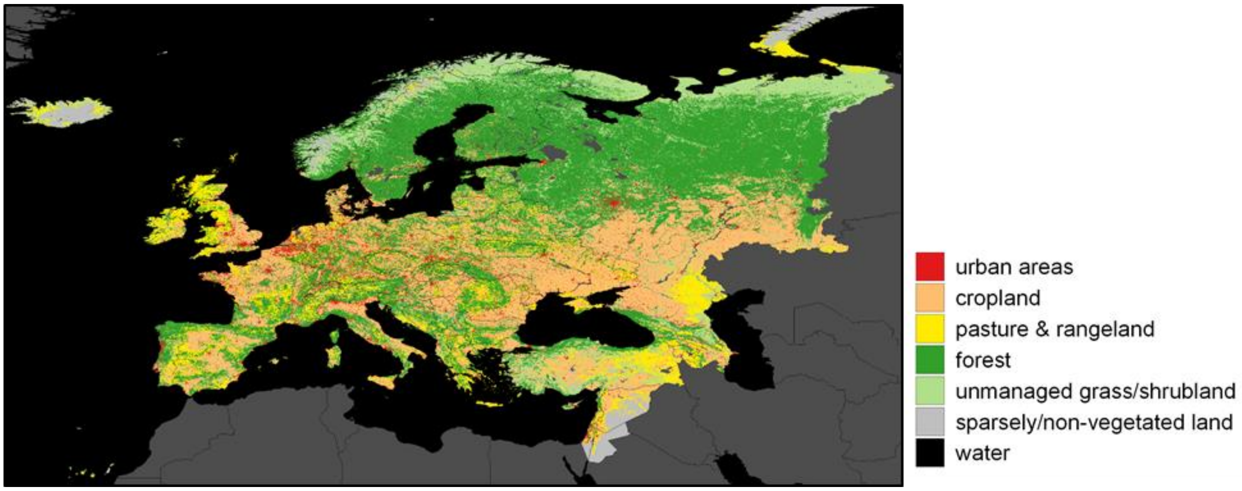

Figure 2: Land use/cover map (quick view) for Europe in 2015 from HILDA+

Updates compared to the previous HILDA+ 0.1 version (vEUR-beta):

- We changed the gross change accounting based on country-specific mean transition matrices from CORINE, ESA-CCI and MODIS Land Cover time series (gross change to class share ratio).
- We included applied 1-yearly time steps for the entire period from 1900 to 2015. Note that observational data is missing before 1960. Thus, we extrapolated the temporal trend of the used probability maps back in time for each land use/cover category. The same was done for the FAO land use database.
- We updated the projection to EPSG:4326/ WGS84 with latitude/longitude coordinates.

In HILDA+ vEUR-0.1, the underlying data basis for the allocation procedure differs depending on the period:

- 1960-2015: reported FAO land use trends and remote sensing-based class probability maps for change allocation (probability maps are the results of harmonising multiple EO-based land cover products with high resolution)
- 1900-1959: extrapolated FAO land use trends and extrapolated class probability maps (no underlying observational or reported data for this period)

HILDA+ vEUR-0.1 has the following specifications:

Two spatial datasets, hildaplus_vEUR-0.1_LULC_states.nc, which contains annual land cover/use types (6 land use/cover classes), and hildaplus_vEUR-0.1_LULC_transitions.nc, which contains annual land cover/use transition types between two years, are provided with respective documentation. Additionally, tabular data with land transition matrices for each country and each time are provided as CSV files (within the folder TransitionTables).

- Data format: netCDF4
- CRS/ projection: +proj=longlat +datum=WGS84 +no_def (EPSG:4326/ WGS84)
- Temporal coverage: 1900-2015
- Spatial coverage: -48.0635857762923351, 7.6448957863115652: 87.9715743535540184, 81.8717912632577764
- Temporal resolution: 1 year

- Spatial resolution: 0.0129978 degree

4.2.2. Biomass maps

There are unprecedented opportunities to provide large area forest biomass data created by a series of active and upcoming space-based missions. Many of them provide open data targeted at large area and better spatial resolution biomass monitoring than has previously been achieved. In particular, the Climate Change Initiative Biomass (CCI Biomass) project of the European Space Agency (ESA) is providing multiple global biomass data and information mainly for climate modelling and assessments.

Part of the VERIFY project is also to look into such new biomass products and their validation and comparison with plot based reference data sources such as national forest inventory (NFI) datasets. As specific work items, the biomass plot to map comparison was first done using the ESA's 2010 GlobBiomass map and we developed a workflow for map comparison that was also implemented to other global maps e.g. Baccini and GEOCARBON, allowing inter-comparison of global biomass maps. A tier-level validation was used for the 2017 CCI Biomass map. The following datasets have been used in the inter-comparison using large database of reference plots: Baccini (year 2000, 30 m resolution), GEOCARBON (Year ~2008, 500 m resolution), Globbiomass (year 2010, 100m resolution), and Biomass-CCI (year 2017, 100 m resolution). These individual datasets are not comparable for change directly but by using multi-temporal plot reference datasets (done at 0.1 deg resolution), we will be able to assess regional bias and aim for a more harmonized, consistent time series covering the years 2000, 2010, 2017. The Biomass-CCI map for 2018 is in preparation and will be available in 2020.
The variance of misrepresentation error and other random plot error sources was modelled and compared to the uncertainty layer of the biomass CCI map. A method to estimate spatial correlation of above ground biomass (AGB) for small plot sizes was tested using data from contiguous sub-plots within large plots and data from LiDAR campaigns. Aside from measurement error, uncertainties from three more plot errors (harmonization, geo-location, and representation) were estimated. From these estimates, plots were weighted and integrated into bias modelling using a model-based approach. Initial bias maps at global scale were developed and discussed with some of climate users project partners (i.e. LCSE).
Beside the bias, the precision will be modelled to complete the spatial error diagnostics of global biomass maps. Once a thorough understanding of the uncertainties of these new biomass maps are provided, their usefulness for national reporting and estimation will be further explored, in particular with the 2019 IPCC refinement of the GHG-I guidelines that includes a new section introducing the use of biomass estimates from maps generated from space-based data.
The use of biomass maps is increasingly important, as these wall-to-wall datasets have the potential to complement plot-based biomass measurements available through NFIs.

- Data format: geotiff
- CRS/ projection: +proj=longlat +datum=WGS84 +no_def
  (EPSG:4326/ WGS84)
- Temporal coverage: 2000, 2010, 2017
- Spatial coverage: -48.0635857762923351, 7.6448957863115652: 87.9715743535540184, 81.8717912632577764
- Temporal resolution: 1 year

- Spatial resolution: 0.1 deg

4.3. Soil property and soil erosion datasets

There are no changes in the provided datasets since the last deliverable (see D3.1).

4.4. Flux datasets for model testing

There are no changes in the provided datasets since the last deliverable (see D3.1).

4.5. Cropland management data

4.5.1. CAPRI fertilizer application data

There are no changes in the provided datasets since the last deliverable (see D3.1).

4.5.2. Management timing

As described in the last deliverable D3.1 the data for sowing and harvest are calculated by phenological models according to Waha et al. (2012) and van Bussel et al. (2015). The models base the calculations on temperature and precipitation data for the different areas. As there was a modification of the climate data (section 2.1.) the timing of the management needed to be re-calculated. The dates for sowing and harvest are calculated for spring wheat and maize in Europe. Further crops will be added to the calculations.

References:

van Bussel, L. G. J., Stehfest, E., Siebert, S., Müller, C., & Ewert, F. (2015). Simulation of the phenological development of wheat and maize at the global scale. Global Ecology and Biogeography, 24(9), 1018–1029. https://doi.org/10.1111/geb.12351

Waha, K., van Bussel, L. G. J., Müller, C., & Bondeau, A. (2012). Climate-driven simulation of global crop sowing dates. Global Ecology and Biogeography, 21(2), 247–259. https://doi.org/10.1111/j.1466-8238.2011.00678.x

4.6. Grassland management data

There are no changes in the provided datasets since the last deliverable (see D3.1).

4.7. Forest management data

There are no changes in the provided datasets since the last deliverable (see D3.1).

4.8. Nitrogen deposition data

There are no changes in the provided datasets since the last deliverable (see D3.1).

4.9. Freshwater fluxes and river exports

There are no changes in the provided datasets since the last deliverable (see D3.1).

4.10. Coastal ocean CO2 fluxes

4.10.1. Introduction

For estimating the air-sea CO_2 gas exchange in European shelf seas, we generated maps of sea surface pCO_2 covering the area from the western Mediterranean to the Barents Sea and then calculated the fluxes based on these maps combined with the atmospheric xCO_2 in the marine boundary layer and 6 hourly wind speed data. The pCO_2 maps were generated by fitting a set of driver data (for example sea surface temperature, mixed layer depth or chlorophyll concentration) against gridded fCO_2 observations from SOCAT (Bakker et al., 2016). These fits were then applied on available maps of driver data to make pCO_2 maps. We used two different fit routines:

- a multi linear regression
- and a random forest approach.

We have produced three different versions of these maps, using different combinations of driver data and fit routine. The maps have a spatial resolution of 0.125x0.125, are available in a monthly version or with a 6-hourly temporal resolution and cover the time span from 1998 to 2018. 2019 is available as a prediction with larger uncertainties than the rest of the dataset.

4.10.2. Methods:

Due to the large oceanographic and biogeochemical variability over the European shelf seas, the region was divided into a set of subregions (Figure 3). This was based on the COastal Segmentation and the related CATchments (COSCAT) segmentation scheme (Laruelle et al., 2013) and also the data coverage in the respective regions. We defined data to be coastal if they were obtained in regions with water depth of less than 500 m or within 100 km from shore.

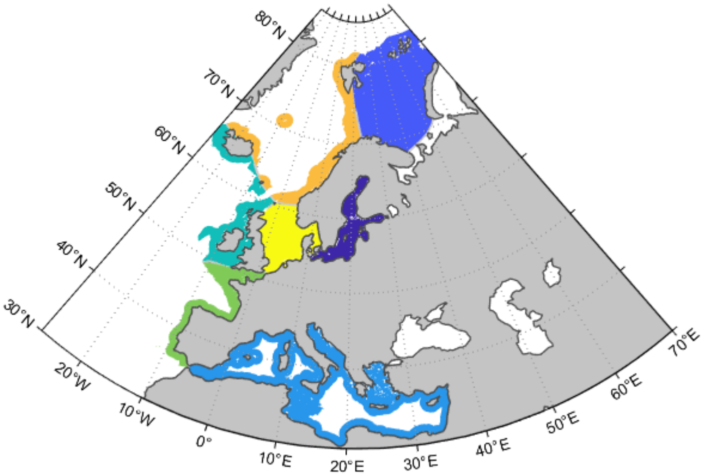

Figure 3: Overview over the study area and the 7 different subregions.

4.10.3. Input data

We extracted available $f\mathrm{CO}_2$ observations on the European shelf from the SOCAT (Bakker et al., 2016) database (www.socat.info, version 2019, quality flags A-E) and gridded these data on a monthly 0.125° x 0.125° grid.

The driver data used for the fitting routines are listed in **Erreur ! Source du renvoi introuvable.**; these are supplied as 3D mapped distributions (lat x lon x time). For establishing the statistical fits, driver data collocated in time and space with the SOCAT $f\mathrm{CO}_2$ observations were extracted from the mapped distributions and gridded on to the monthly 0.125° x 0.125° grid. This ensures that the driver data used for the statistical fitting has spatial and temporal coverage corresponding to the $f\mathrm{CO}_2$ observations. For producing the maps, the driver data were regridded to the monthly 0.125° x 0.125° grid, providing the full spatial and temporal coverage and a homogenous average in each grid box. If a different resolution of the respective products were used for producing fit than for producing the actual map, these are denoted with fit and model in Table 2. For a detailed description see (Becker et al., 2020).

Product used	Resolution	Reference
$f\mathrm{CO}_2$ observations		SOCAT dataset (Bakker et al., 2016)
Chl *a* (fit)	4 km x 4 km, 8 days	Global Ocean Chlorophyll II (Copernicus-GlobColour) from Satellite Observations - Reprocessed
Chl *a* (model)	4 km x 4 km, monthly	Global Ocean Chlorophyll II (Copernicus-GlobColour) from Satellite Observations - Reprocessed
MLD/SST/SSS (fit)	0.25° x 0.25°, weekly	(Guinehut et al., 2012)
MLD/SST/SSS (model)	0.25° x 0.25°, monthly	(Guinehut et al., 2012)
BATHYMETRY	2 min x 2min	(National Geophysical Data Center, 2006)
ICE	0.25° x 0.25°, monthly	(Cavalieri et al., 1996)
***x*CO_2, atmosphere**	10 zonal, latitudinal bands, monthly	NOAA Greenhouse Gas Marine Boundary Layer Reference

Rödenbeck *p*CO₂	5˚ x 4˚, monthly	(Rödenbeck et al., 2014)
Wind speed	6-hourly	(Kanamitsu et al., 2002)

Table 2 : Products used as driver data and for calculating the fluxes

4.10.4. Multi linear regression (MLR)

The multi linear regression models were constructed by forward and backward stepwise regression using the driver data as predictor variables to model the *f*CO$_2$ observations. In a stepwise regression, the impact of adding or removing a variable from the set of predictor variables in the MLR is tested. The decision on whether to add or remove a variable is based on the p-value of the F-statistic of the MLR. The entrance tolerance was set to 0.05 and the exit tolerance to 0.1. The model includes constant, linear, and quadratic terms as well as products of linear terms. Equation (1à gives the basic equation, with $X_1...X_n$ being the driver data and $a_1...a_{nn}$ the regression coefficients and *y* is the *f*CO$_2$.

$$y = a_0 + a_1 \cdot X_1 + ... + a_n \cdot X_n + a_{12} \cdot X_1 X_2 + ... + a_{mn} \cdot X_m X_n + a_{11} \cdot X_1^2 + ... + a_{nn} \cdot X_n^2 \qquad (1)$$

4.10.5. Random Forest

We use a bagged regression tree model for the random forest fits (Belgiu and Drăguţ, 2016). The tree is built by splitting a random subset of the input dataset into subsets, based on the characteristics of these subsets. We used a number of 500 independent regression trees, each based on a random subset of the input data (leaf size: 20), of which the output then was averaged to obtain the final model response. As far as we are aware, this is the first time Random Forest machine learning is used for pCO$_2$ mapping.

4.10.6. Flux calculation

The air-sea disequilibrium was calculated as the difference between our mapped *f*CO$_2$ values and the atmospheric *f*CO$_2$ in each grid cell and time step. The atmospheric *f*CO$_2$ was determined by converting the *x*CO2 from the NOAA Marine Boundary Layer Reference product from the NOAA GMD Carbon Cycle Group into *f*CO$_2$ by using the monthly SST and SSS data (Table 2) and monthly air pressure data from the NCEP-DOE Reanalysis 2 (Kanamitsu et al., 2002) following the set of equations provided in Pierrot et al. (2009).

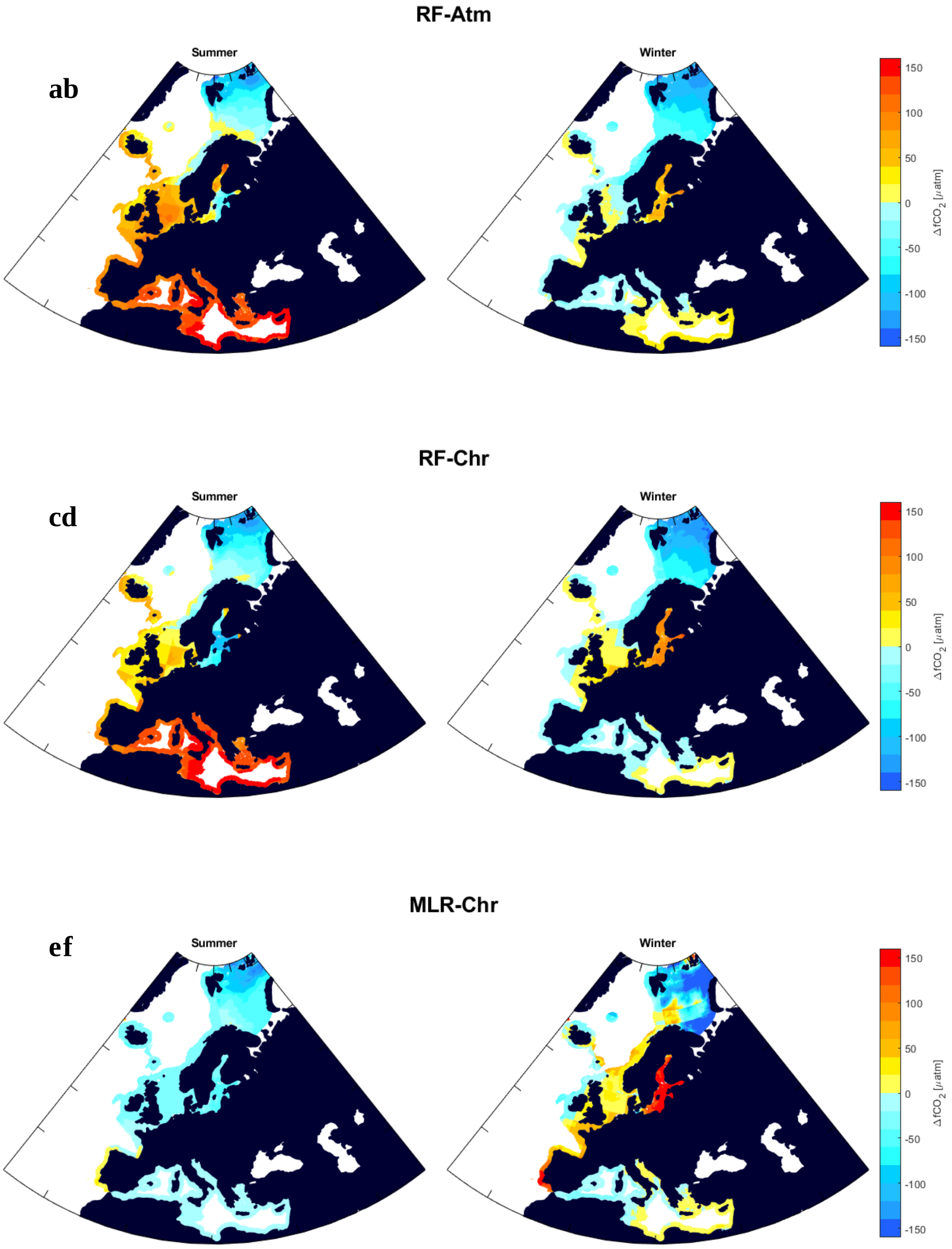

Figure 4: Average air-sea disequilibrium during summer (a, c, e) and winter (b. d. f) as an example for all three products.

We calculated the air-sea CO2 flux (*F*) according to Equation (2), such that negative fluxes are into the ocean. The gas transfer velocity, *k*, was determined using the quadratic wind speed (*u*) dependency of (Wanninkhof, 2014) (Equation (3)). The Schmidt number, Sc, was calculated according to of (Wanninkhof, 2014) and the solubility coefficient for CO_2, K_0, after (Weiss, 1974).

$$\mathrm{F} = k \cdot K_0 \cdot (f\mathrm{CO}_{2,\mathrm{sw}} - f\mathrm{CO}_{2,\mathrm{atm}}) \qquad (2)$$

$$k = a_q \cdot \langle u^2 \rangle \cdot \left(\frac{Sc}{660}\right)^{-0.5} \qquad (3)$$

In our calculations, we used 6-hourly winds of the NCEP-DOE Reanalysis 2 product. The coefficient a_q in Equation 3 is strongly dependent on the wind product used. We determined it to be a_q = 0.16 cm h^{-1} for the 6-hourly NCEP 2 product following the recommendations of (Naegler, 2009) and by using the World Ocean Atlas sea surface temperatures (Locarnini et al., 2018). For the monthly product the monthly mean of the second moment of the NCEP2 6-hourly wind speeds was used to determine *k*.

Presence of sea ice acts as a barrier on the flux. This effect was taken into account by reducing the flux in proportion to the ice cover whenever present, following Loose et al. (2009). As the gas exchange in areas that are considered 100% ice covered from satellite images should not be completely neglected, we used a sea ice barrier effect for a 99% sea ice cover in all grid cells where the sea ice coverage exceeded 99%.

4.10.7. Dataset description

We produced three different maps based on different combinations of driver data and regression model (for all, the fluxes are available at monthly and a 6-hourly resolutions):

1 **RF-Atm:** random forest regression, driver data: SST, MLD, SSS, Chl *a*, BAT, ICE, Lon, Lat, $xCO_{2(atm)}$

2 **RF-Chr:** random forest regression, driver data: SST, MLD, SSS, Chl *a*, BAT, ICE, open ocean pCO_2 (Rödenbeck et al., 2014)

3 **MLR-Chr:** multi linear regression, driver data: SST, MLD, SSS, Chl *a*, BAT, ICE, open ocean pCO_2 (Rödenbeck et al., 2014)

There are three different NetCDF files available:

1. The pCO_2 of the three different models (0.125 x 0.125˚, monthly, 1998-2019)
2. The monthly air-sea CO_2 flux determined from the three different models (0.125 x 0.125˚, monthly, 1998-2019)
3. The air-sea-CO_2 flux in full resolution (6-hourly) determined from the three different models (0.125 x 0.125˚, 6-hourly, 1998-2019)
   (for this product contact meike.becker@uib.no)

Figure 4 shows average air-sea disequilibrium during winter and summer as an example for all three products. Statistical measures for the three models can be found in Table 3. This includes the root mean square error of the fit (RMSE), the Nash Sutcliffe Method Efficiency (ME) (Nondal et al., 2009), and $R^{Seasonal}$ is the relative monthly mismatch, calculated after:

$$R^{Seasonal} = \frac{M}{M_{bench}} \times 100$$

with M being the standard deviation of the monthly model data minus a monthly average over all observations in that biome. M_{bench} is used to relate the amplitude to the usual variability in the region by using the annual mean in reach region instead of the monthly model data.
The three models cover the time range from 1998 to 2018 with the here shown uncertainty. The model output for 2019 has a somewhat higher uncertainty for two reasons: (1) the $f\text{CO}_2$ observations that were used to train the models only reach until December 2018 and (2) some of the reanalysis products that were used to produce the models were not available for 2019 (for example the SST/SSS/MLD product). Here we used the near-realtime products for the year 2019.

All these datasets are first available to the VERIFY project partners and will be freely available through the VERIFY database (see section 3) at the end of 2020.

Model	**Subregion**						
	Barents Sea	Baltic Sea	Mediterranean	Great Britain	Atlantic Coast	Nordic Seas	North Sea
RF-ATM							
RMSE	9.5	36.0	9.5	23.0	21.0	18.0	21.5
ME	0.10	0.36	0.11	0.22	0.43	0.03	0.50
$R^{Seasonal}$	28	37	52	56	46	61	86
RF-Chr							
RMSE	9.0	24.0	8.5	20.5	15.0	15.5	17.0
ME	0.05	0.09	0.01	0.05	0.10	0.09	0.25
$R^{Seasonal}$	29	21	48	44	29	47	70
MLR-Chr							
RMSE	15.1	39.2	12.8	26.7	23.4	22.1	26.5
ME	0.4	3.5	0.6	0.1	5.0	1.8	0.6
$R^{Seasonal}$	66	59	50	58	104	62	70

Table 3 : Statistical measures of the three different models.

References:

Bakker, D.C.E., Pfeil, B., Landa, C.S., Metzl, N., O'Brien, K.M., Olsen, A., Smith, K., Cosca, C., Harasawa, S., Jones, S.D., Nakaoka, S., Nojiri, Y., Schuster, U., Steinhoff, T., Sweeney, C., Takahashi, T., Tilbrook, B., Wada, C., Wanninkhof, R., Alin, S.R., Balestrini, C.F., Barbero, L., Bates, N.R., Bianchi, A.A., Bonou, F., Boutin, J., Bozec, Y., Burger, E.F., Cai, W.-J., Castle, R.D., Chen, L., Chierici, M., Currie, K., Evans, W., Featherstone, C., Feely, R.A., Fransson, A., Goyet, C., Greenwood, N., Gregor, L., Hankin, S., Hardman-Mountford, N.J., Harlay, J., Hauck, J., Hoppema, M., Humphreys, M.P., Hunt, C.W., Huss, B., Ibánhez, J.S.P., Johannessen, T., Keeling, R., Kitidis, V., Körtzinger, A., Kozyr, A., Krasakopoulou, E., Kuwata, A., Landschützer, P., Lauvset, S.K., Lefèvre, N., Monaco, C.L., Manke, A., Mathis, J.T., Merlivat, L., Millero, F.J., Monteiro, P.M.S., Munro, D.R., Murata, A., Newberger, T., Omar, A.M., Ono, T., Paterson, K., Pearce, D., Pierrot, D., Robbins, L.L., Saito, S., Salisbury, J., Schlitzer, R., Schneider, B., Schweitzer, R., Sieger, R., Skjelvan, I., Sullivan, K.F., Sutherland, S.C., Sutton, A.J., Tadokoro, K., Telszewski, M., Tuma, M., Heuven, S.M.A.C. van, Vandemark, D., Ward, B., Watson, A.J., Xu, S., 2016. A multi-decade record of high-quality fCO_2 data in version 3 of the Surface Ocean CO_2 Atlas (SOCAT). Earth System Science Data 8, 383–413. https://doi.org/10.5194/essd-8-383-2016

Becker, M., Olsen, A., Landschützer, P., Omar, A., Rehder, G., Rödenbeck, C., Skjelvan, I., 2020. The northern European shelf as increasing net sink for CO_2. Biogeosciences Discussions 1–28. https://doi.org/10.5194/bg-2019-480

Belgiu, M., Drăguţ, L., 2016. Random forest in remote sensing: A review of applications and future directions. ISPRS Journal of Photogrammetry and Remote Sensing 114, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Cavalieri, D.J., Parkinson, C.L., Gloersen, P., Zwally, H.J., 1996. Sea Ice Concentrations from Nimbus-7 SMMR and DMSP SSM/I-SSMIS Passive Microwave Data, Version 1, Monthly.
Guinehut, S., Dhomps, A.-L., Larnicol, G., Traon, P.-Y.L., 2012. High resolution 3-D temperature and salinity fields derived from in situ and satellite observations. Ocean Science 8, 845–857. https://doi.org/10.5194/os-8-845-2012

Kanamitsu, M., Ebisuzaki, W., Woollen, J., Yang, S.-K., Hnilo, J.J., Fiorino, M., Potter, G.L., 2002. NCEP–DOE AMIP-II Reanalysis (R-2). Bull. Amer. Meteor. Soc. 83, 1631–1644. https://doi.org/10.1175/BAMS-83-11-1631

Laruelle, G.G., Dürr, H.H., Lauerwald, R., Hartmann, J., Slomp, C.P., Goossens, N., Regnier, P. a. G., 2013. Global multi-scale segmentation of continental and coastal waters from the watersheds to the continental margins. Hydrology and Earth System Sciences 17, 2029–2051. https://doi.org/10.5194/hess-17-2029-2013

Locarnini, R.A., Mishonov, A.V., Baranova, O.K., Boyer, T.P., Zweng, M.M., GarciA, H.E., Reagan, J.R., Seidov, D., Weathers, K., Paver, C.R., Smoylar, I., 2018. World Ocean Atlas 2018, Volume 1: Temperature. A. Mishonov Technical Ed.

Loose, B., McGillis, W.R., Schlosser, P., Perovich, D., Takahashi, T., 2009. Effects of freezing, growth, and ice cover on gas transport processes in laboratory seawater experiments. Geophysical Research Letters 36. https://doi.org/10.1029/2008GL036318

Naegler, T., 2009. Reconciliation of excess 14C-constrained global CO2 piston velocity estimates. Tellus B: Chemical and Physical Meteorology 61, 372–384. https://doi.org/10.1111/j.1600-0889.2008.00408.x

National Geophysical Data Center, 2006. 2-minute Gridded Global Relief Data (ETOPO2) v2. NOAA. https://doi.org/doi:10.7289/V5J1012Q

Nondal, G., Bellerby, R.G.J., Oldenc, A., Johannessena, T., Olafssond, J., 2009. Optimal evaluation of the surface ocean CO2 system in the northern North Atlantic using data from voluntary observing ships.

Limnol. Oceanogr. 7, 109–118.

Rödenbeck, C., Bakker, D.C.E., Metzl, N., Olsen, A., Sabine, C., Cassar, N., Reum, F., Keeling, R.F., Heimann, M., 2014. Interannual sea–air CO_2 flux variability from an observation-driven ocean mixed-layer scheme. Biogeosciences 11, 4599–4613. https://doi.org/10.5194/bg-11-4599-2014
Wanninkhof, R., 2014. Relationship between wind speed and gas exchange over the ocean revisited: Gas exchange and wind speed over the ocean. Limnology and Oceanography: Methods 12, 351–362. https://doi.org/10.4319/lom.2014.12.351

Weiss, R.F., 1974. Carbon dioxide in water and seawater: The solubility of a non-ideal gas. Mar. Chem. 2, 203–215.

4.10.8. XCO2 from OCO-2 via FOCAL algorithm

The fast atmospheric trace gas retrieval for OCO2 (FOCAL-OCO2) has been setup to retrieve XCO2 (the column-average dry-air mole fraction of atmospheric CO2) by analyzing hyper spectral solar backscattered radiance measurements of NASA's OCO2 satellite. FOCAL includes a radiative transfer model, which has been developed to approximate light scattering effects by multiple scattering at an optically thin scattering layer. This reduces the computational costs by several orders of magnitude. FOCAL's radiative transfer model is utilized to simulate the radiance in all three OCO-2 spectral bands allowing the simultaneous retrieval of CO2, H2O, and solar induced chlorophyll fluorescence.

The FOCAL OCO-2 XCO2 version 08 data product has been extended by two years and now covers the time period 2015-2018. Data access and other information is provided on the FOCAL-OCO2 website (http://www.iup.uni-bremen.de/~mreuter/focal.php). Given the size of the data stream we did not copy this data set on the VERIFY THREDDS server.

5. Organization of the database

A list of available data and the data itself are available from the VERIFY THREDDS data server (TDS, https://verifydb.lsce.ipsl.fr/thredds/verify/catalog.html), with some limited metadata available from the VERIFY data catalogue (available from the VERIFY web site: http://verify.lsce.ipsl.fr/index.php/products). Note that for the VERIFY partners additional information on the different datasets is available under the password protected share point platform (https://projectsworkspace.eu/sites/VERIFY/Lists/WP3inputdataset/AllItems.aspx).

The filenames will be assigned to contain various information about the file itself, including the method, species, institute, region, spatial coverage, temporal resolution, and the person who uploaded the file. This information is used to automatically generate a catalogue of available data (http://webportals.ipsl.jussieu.fr/VERIFY/CountryTot2.html). The TDS is developed by Unidata, a member of the UCAR Community Programs, managed by the University Corporation for Atmospheric Research, and funded by the National Science Foundation of the United States, with the goal of helping educators and scientists obtain and use geoscience data. The TDS also supports several dataset collection services including some sophisticated dataset aggregation capabilities. This allows the TDS to aggregate a collection of datasets into a single virtual dataset, greatly simplifying user access to that data collection. The TDS also contains viewing tools to facilitate direct user browsing of stored datasets, instead of forcing the user to rely on metadata.

More details are available in Deliverable 6.8.

6. Conclusions

The VERIFY database is a living structure, growing out of collaboration among various workpackages (primarily WP3 and WP4). As such, both it and this document will continue to be updated as user needs evolve. Part of the work currently being done in VERIFY is to determine the most efficient and scientifically accurate way to share datastreams among the various research groups. Such work is reflected in the VERIFY database, which consists of these datastreams. Some of the datastreams have not yet been fully taken advantage of, but this will be a focus of the next phase of the project.