ENVIRONMENTAL RESEARCH
LETTERS

OPEN ACCESS

RECEIVED
25 March 2022

REVISED
26 September 2022

ACCEPTED FOR PUBLICATION
24 October 2022

PUBLISHED
4 November 2022

Original content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence.

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

LETTER

Archetypes of agri-environmental potential: a multi-scale typology for spatial stratification and upscaling in Europe

Michael Beckmann[1,6], Gregor Didenko[1,6], James M Bullock[2], Anna F Cord[3], Anne Paulus[1], Guy Ziv[4] and Tomáš Václavík[5,*]

[1] UFZ—Helmholtz Centre for Environmental Research, Department of Computational Landscape Ecology, Permoserstr. 15, 04318 Leipzig, Germany
[2] UK Centre for Ecology & Hydrology, Benson Lane, Wallingford, Oxfordshire OX10 8BB, United Kingdom
[3] Chair of Computational Landscape Ecology, Institute of Geography, Technische Universität Dresden, Helmholtzstr. 10, 10169 Dresden, Germany
[4] University of Leeds, School of Geography, Faculty of Environment, Leeds LS2 9JT, United Kingdom
[5] Palacký University Olomouc, Faculty of Science, Department of Ecology and Environmental Sciences, Šlechtitelů 27, 78371 Olomouc, Czech Republic
[6] Shared first authorship.
[*] Author to whom any correspondence should be addressed.

E-mail: tomas.vaclavik@upol.cz

Keywords: archetypes, multi-scale, typology, land systems, self-organizing maps

Abstract

Developing spatially-targeted policies for farmland in the European Union (EU) requires synthesized, spatially-explicit knowledge of agricultural systems and their environmental conditions. Such synthesis needs to be flexible and scalable in a way that allows the generalization of European landscapes and their agricultural potential into spatial units that are informative at any given resolution and extent. In recent years, typologies of agricultural lands have been substantially improved, however, agriculturally relevant aspects have yet to be included. We here provide a spatial classification approach for identifying archetypal patterns of agri-environmental potential in Europe based on machine-learning clustering of 17 variables on bioclimatic conditions, soil characteristics and topographical parameters. We improve existing typologies by (a) including more recent biophysical data (e.g. agriculturally-important soil parameters), (b) employing a fully data-driven approach that reduces subjectivity in identifying archetypal patterns, and (c) providing a scalable approach suitable both for the entire European continent as well as smaller geographical extents. We demonstrate the utility and scalability of our typology by comparing the archetypes with independent data on cropland cover and field size at the European scale and in three regional case studies in Germany, Czechia and Spain. The resulting archetypes can be used to support spatial stratification, upscaling and designation of more spatially-targeted agricultural policies, such as those in the context of the EU's Common Agricultural Policy post-2020.

1. Introduction

Current land management dynamics are driven by social, economic and political changes (Stoate *et al* 2009, Batáry *et al* 2015, Lomba *et al* 2015), which are putting European agroecosystems under an immense pressure and leading to land-use intensification (to achieve higher cost-effectiveness) in some areas and land abandonment in others (Plieninger *et al* 2016). Given that nearly half of the land in the European Union (EU) is used for agriculture (Castillo *et al* 2018), sustainable management of agro-ecosystems is key to preventing further degradation of farmland and the ecosystem services they provide, and to achieving the EU's environmental and climate objectives (Rega *et al* 2020). At the same time, Europe's agricultural landscapes are highly diverse due to climatic, biophysical and socioeconomic differences typically encountered at continental scales. Past agricultural policies have commonly been criticized for oversimplifying the inherent complexity of Europe's agricultural systems and trying to impose

© 2022 The Author(s). Published by IOP Publishing Ltd

one-size-fits-all solutions for subsidization and regulation of the EU's agricultural sector (e.g. Bureau *et al* 2012, PBL 2012). Indeed, policies and actions have different outcomes depending on the type of agricultural system targeted, the type of farming and land use, or local socio-economics (Ziv *et al* 2020). Developing policies that overcome these shortcomings and are tailored to fit national, regional or even local scales could be supported by spatially-explicit typologies that capture archetypal patterns of agri-environmental systems. Such typologies need to be flexible and scalable in a way that allows the classification of agricultural landscapes into spatial units that are informative at any given scale and extent.

Great efforts have been devoted in recent years to developing methods to identify and map archetypal patterns of agricultural systems, particularly in Europe (e.g. Andersen 2017, Levers *et al* 2018, Rega *et al* 2020). In addition to such continental-scale archetypes, others have mapped land system archetypes at smaller scales (e.g. Janík and Romportl 2016, Malek and Verburg 2017, Dittrich *et al* 2019). However, most of these archetypes have been prepared for specific applications (e.g. for mapping crop-management systems, exploring changes in land-use intensities, or understanding bundles of ecosystem services), often relying on data that are difficult to obtain and share (e.g. census data on individual crops or data from the Farm Accountancy Data Network). In contrast, more general characterizations of agricultural landscapes, that rely mostly on biophysical factors such as topography, climate, or land cover, have proved to be highly useful for upscaling of regional findings across the continent, for the selection of representative case studies, or as frameworks for modeling land use and policy impacts (Hazeu *et al* 2010, Mücher *et al* 2010, Metzger *et al* 2013, Václavík *et al* 2016). We here aim to bridge these approaches by providing a novel and freely accessible base map of agri-environmental potential in Europe, which can be adapted and scaled to fit the requirements of other study contexts (e.g. socio-economic studies, behavioral studies, species distribution modeling).

We present a spatial classification approach for identifying archetypal patterns of agri-environmental potential in Europe. We define archetypes as recurrent patterns in variables and processes that shape land and social-ecological systems and can be expressed as typologies of cases (sensu Oberlack *et al* 2019). In order to support spatial targeting of agricultural policies, upscaling and transferability of regional findings and other application domains (figure 1), we provide a development beyond existing typologies by (a) including more recent biophysical data that have become available, including agriculturally-important soil parameters which have not been included in previous archetypal classifications (Hengl *et al* 2017), (b) employing a fully data-driven approach to define rules for creating archetypes, which allows more flexibility when adapting the archetypes to specific study requirements and (c) providing an easy way to adjust the archetypes by defining the number of spatial clusters that allows scalable results suitable both for the entirety of Europe as well as smaller geographical extents and fits best to the specific study purpose. We demonstrate the utility, flexibility and scalability of our approach by comparing the archetypes with independent data on cropland cover and field size at the continental (European) scale and at the regional scale in three regional case studies in Germany, Czechia and Spain. The resulting archetypes can be used to support decision-making and designation of more spatially targeted agricultural policies, especially in the context of the EU Common Agricultural Policy post-2020 and the EU Biodiversity Strategy towards 2030.

2. Methods

2.1. Data and variable selection

This study's extent is Europe, covering approximately 6.63 million km^2 (figure 2). The input datasets (table 1) that we used to identify agri-environmental archetypes were chosen to cover the biophysical variation of agri-environmental systems in Europe, especially in terms of climate, soil and topographical parameters. However, we did not restrict our analysis to agricultural land only. We adopted a broader view of agri-environmental archetypes, referring to them as spatial units with similar biophysical characteristics related to land suitability and potential agricultural production. Our variables reflect the basic determinants of modern agricultural production capacity, similarly as in the case the Agro-Ecological Zones (Hazeu *et al* 2010), controlling what agricultural systems have the potential to be in a certain location in the absence of human decisions, political history, market structures, implementation of the Common Agricultural Policy, etc.

First, we included 19 bioclimatic variables from the WorldClim database v2 (Fick and Hijmans 2017; www.worldclim.org), which contains long-term global climate and bioclimatic variables at 1 km resolution. Bioclimatic indicators provide a useful basis for environmental stratification. They describe seasonal conditions and climate extremes and, thus, they are considered to be more agriculturally relevant than monthly climate observation (Galdies and Vella 2019). To include a variable reflecting the length of agricultural production, we calculated growing degree days (GDD) using the summed temperature of all months with an average temperature higher than 5 °C multiplied by the number of days.

Soil properties are important determinants of farming systems, and so, second, we acquired the

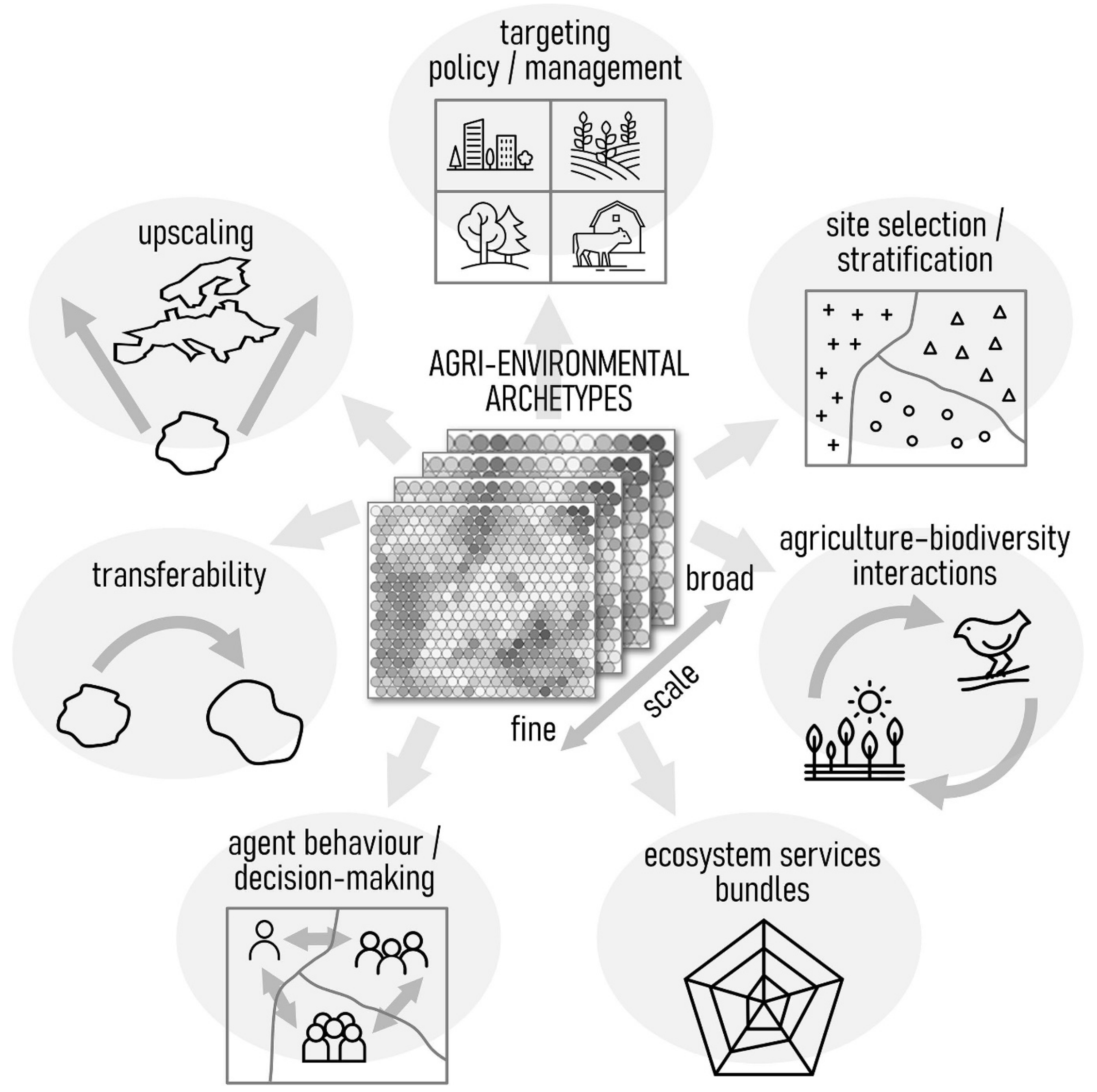

Figure 1. Conceptual framework showing the scalable approach for identifying archetypes of agri-environmental potential and their application domains. These include spatial tailoring of policies to fit national or regional needs; stratifying regions for selection of research sites or assessing geographical representativeness; providing a modeling framework for investigating the interactions between farming and biodiversity in agro-ecosystems, for assessing bundles of ecosystem services, or for modeling the complexity of decision making and behavior of agents in the agricultural sector; assessing the transferability of place-based research to other regions with similar agri-environmental characteristics; and upscaling of land-use models and management recommendations developed in regional case studies to larger geographical extents.

SoilGrids database of 15 global gridded and harmonized soil variables at the 250 m resolution (Hengl *et al* 2017; www.isric.org/explore/soilgrids). We selected a soil depth of 30 cm (most relevant for farming) and transformed all raster datasets to Lambert azimuthal equal area projection, warping them with bilinear resampling warp method to a resolution of 1 km to form a spatially consistent basis of input data. Third, topographic variation underlies most patterns and processes in land systems and is key to understanding spatial variation in land use and agricultural activities. To express the main topographical characteristics, we extracted elevation and terrain ruggedness index (TRI) from the Global Multi-resolution Terrain Elevation Data (GMTED) available from the EarthEnv database (Amatulli *et al* 2018) at a 1 km resolution.

To avoid collinearity and redundancy in the input information, we inspected Pearson correlation coefficients between all variables (figures A1–A3), using $r = |0.7|$ as a conservative threshold of collinearity (Dormann *et al* 2013). If two variables were correlated, only one was kept for further analysis, giving preferences to those with more direct agro-environmental relevance. For example, GDD was highly correlated with Annual mean temperature, therefore only the temperature variable was retained. However, we made two exceptions: (a) sand, clay and silt are the building blocks of soil, therefore correlated but still individually important for agriculture; and

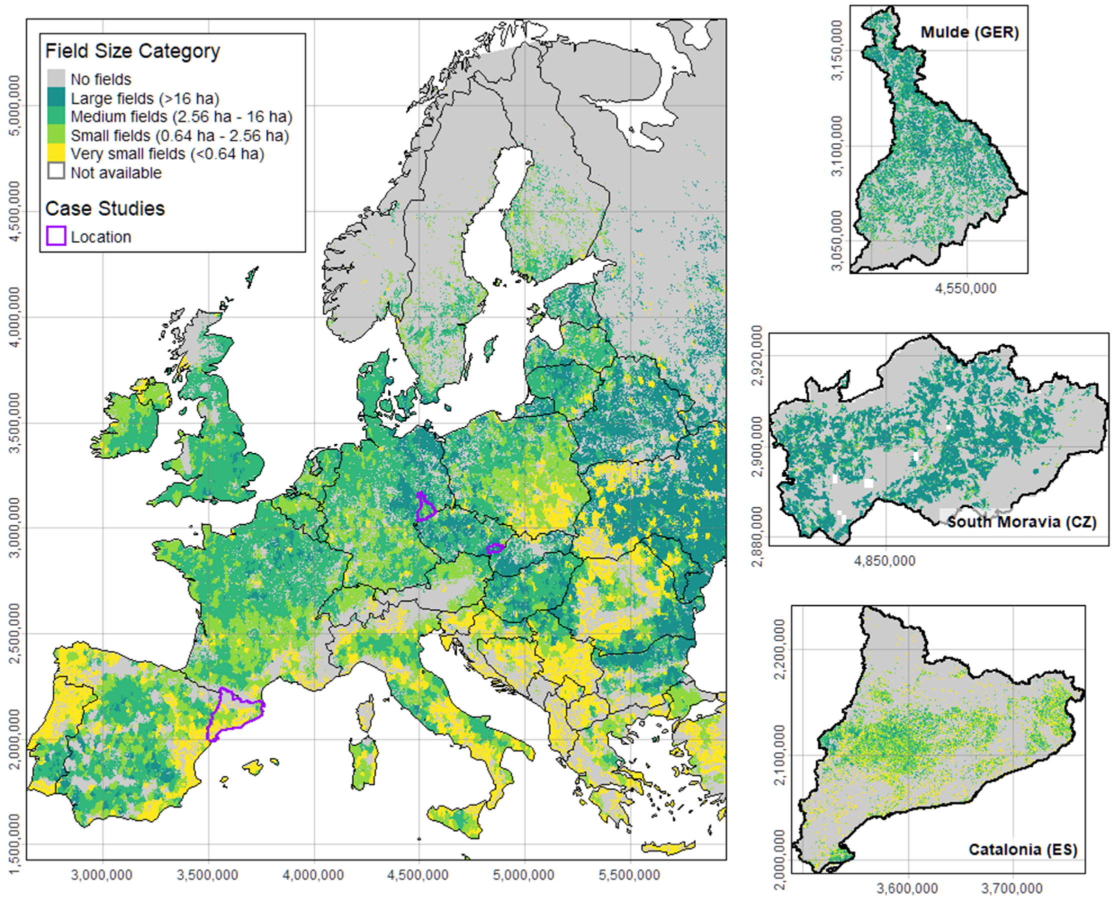

Figure 2. The main study area (Europe) and three regional case studies in Germany, Czechia and Spain. The case study regions are shown in purple on the main map and in the map insets on the right. The maps show field sizes (Lesiv *et al* 2019) with categories defined as: very large (>100 ha), large (16 ha–100 ha), medium (2.56 ha–16 ha), small (0.64 ha–2.56 ha) and very small (<0.64 ha) following the GEOGLAM framework. Maps in Lambert azimuthal equal area projection.

(b) elevation and TRI, which had the correlation coefficient slightly above 0.7 but express different characteristics of topography. Our final set of input indicators included 17 variables (table 1). Only cells that had no missing values were used for further analysis. Therefore, 4% of cells (267 184 km^2) were removed, scattered mostly over Scandinavia and the Alps, where no soil information was available.

2.2. Spatial classification of agri-environmental archetypes

We used self-organizing maps (SOMs; Kohonen 2001) to cluster the selected multi-dimensional data into archetypal patterns of agri-environmental potential. SOMs are based on artificial neural networks following a competitive learning algorithm with an input layer (input variables) and an output layer (clusters). The method allows visualization of complex data by reducing their dimensionality to a predefined two-dimensional output space (map) of *k* neurons (or nodes), clustering observations (e.g. grid cells) based on their similarity. SOMs are becoming a common approach for identifying archetypes as typologies of cases (Sietz *et al* 2019) and have been used in several recent studies mapping archetypes of land and social-ecological systems (Václavík *et al* 2013, Levers *et al* 2018, Dittrich *et al* 2019).

First, since the input data had to be standardized to allow for a relatively equal influence of weight vectors (Kohonen 2001), we used z-score normalization to scale all variables to zero mean and standard deviation of 1. Then, we determined the size of the two-dimensional output space. This size is selected prior the classification procedure, with a small number of output nodes forcing the SOM to behave solely as a clustering technique, and a very large number of nodes (exceeding the number of input observations) enabling the emergence of fine-scale patterns (Delmelle *et al* 2013). To assess the utility of our approach at multiple scales, we aimed to find an appropriate number of *k* clusters for both regional and continental applications. Using the heuristic equation approach (Vesanto and Alhoniemi 2000) with a two-stage clustering method (Park *et al* 2014, Cracknell *et al* 2015, Li *et al* 2019) on a very large SOM (112 by 112) showed that the Within-cluster sum of squares (WCSS) of the resulting second-stage Hierarchical agglomerative clustering (HAC) was optimal for 20 clusters (figure A4).

Table 1. Final selection of 17 variables used for classification of agri-environmental archetypes: seven climate derived from WorldClim v2 (Fick and Hijmans 2017), eight soil-related variables originating from the SoilGrids dataset (Hengl *et al* 2017), and two topographic parameters derived from the Global Multi-resolution Terrain Elevation Data (Amatulli *et al* 2018). All variables collated at 1 km resolution.

Origin	Variable name (unit)	Description
Worldclim	Annual precipitation (mm)	Annual sum of precipitation
Worldclim	Precipitation warmest quarter (mm)	Precipitation variable that corresponds with the peak solar radiation and highest growth potential
Worldclim	Mean Diurnal temp. range (°C)	Mean of monthly temperature ranges, providing information on temperature fluctuation
Worldclim	Annual mean temp (°C)	Basic temperature variable with influence on crop cultivation
Worldclim	Annual temp. range (°C)	Displays the range of extreme temperature conditions and serves as a proxy for continentality/oceanicity
Worldclim	Mean temp. wettest quarter (°C)	Basic temperature variable in the wettest quarter of the year
Worldclim	Precipitation seasonality (%)	Variation in monthly precipitation over the course of the year
SoilGrids	Coarse Fragments (%)	Volumetric fraction of coarse fragments
SoilGrids	SOC concentration (g kg^{-1})	Soil organic carbon concentration
SoilGrids	Sand content (%)	Proportion of sand particles (>0.05 mm) in the fine earth fraction
SoilGrids	Bulk density (kg m^{-3})	Bulk density of the fine earth fraction
SoilGrids	PH KCl (index *10)	Soil pH
SoilGrids	Clay content (%)	Proportion of clay particles (<0.002 mm) in the fine earth fraction
SoilGrids	Cation exchange capacity (cmol kg^{-1})	Cation exchange capacity of the soil
SoilGrids	Silt content (%)	Proportion of silt particles (⩾0.002 mm and ⩽0.05 mm) in the fine earth fraction
GMTED	Elevation (m)	Mean elevation in meters above sea level
GMTED	Terrain Ruggedness (m)	Sum of absolute change in elevation between the grid cell and its eight neighbors

To find a number of clusters for the regional application, we investigated the quantization error (QE) of differently sized SOMs, from $k = 9$ to $k = 2500$. QE is a quality measure of the classification procedure, calculated as the distance of each observation to the cluster centroid. It indicates how homogeneous the clusters are: good classifications should show relatively small distances for most observations. We selected $k = 400$ as the optimal number of clusters using the 'Elbow Method' (Kassambara 2017, figure A5).

We used the Geospatial Data Abstraction Library 3.0.2 (GDAL/OGR contributors 2019) for the preparation of all input variables. All other processing and visualization were done in the statistical programming language R 3.5.0 (R Core Team 2019). SOM clustering was implemented with the kohonen 3.0.1 package (Wehrens and Kruisselbrink 2018).

2.3. Comparison to agricultural data

To demonstrate the utility of our typology we compared the outcomes of the SOM clusters with independent data on mean cropland cover and field size. We assumed that identified agri-environmental archetypes, despite not being restricted to agricultural land, should reflect the biophysical conditions that drive some of the variation in agricultural data, e.g. locations with high TRI and temperature extremes co-occurring with small field sizes. At the same time, we assumed that individual input datasets would not be significantly associated with agricultural data; we tested this assumption by calculating Pearson's correlation coefficients between each input variable and cropland cover and field size, respectively, using 1% of randomly selected pixels to avoid spatial autocorrelation (table A1).

To compare the *k*20 clustering approach, we used the global maps of mean cropland cover and agricultural field size developed by International Institute for Applied Systems Analysis-International Food Policy Research Institute (IIASA-IFPRI) at 1 km resolution (figure A6, Fritz *et al* 2015). The product defines cropland as the sum of arable land and permanent crops, following the definition of the Food and Agriculture Organization. The field size map (figure 2) from Lesiv *et al* (2019) defines field size categories as: large (>16 ha), medium (2.56 ha–16 ha), small (0.64 ha–2.56 ha) and very small (<0.64 ha). In this dataset, a field is defined as an enclosed agricultural area, including annual and perennial crops, hayfields and fallow but, in contrast to the cropland cover product, also permanent pastures.

To test the applicability of the *k*400 clustering approach, we acquired data on cropland cover and field size for three case study regions that are part of the European Commission-funded research project BESTMAP (Ziv *et al* 2020): the Saxonian part of the Mulde river basin in Germany, the South Moravia region in the south-eastern part of Czechia, and Catalonia in Spain (figure 2). For the German case study, we obtained field parcel geometries from the InVeKoS database of Saxony (InVeKoS Sachsen—SMEKUL

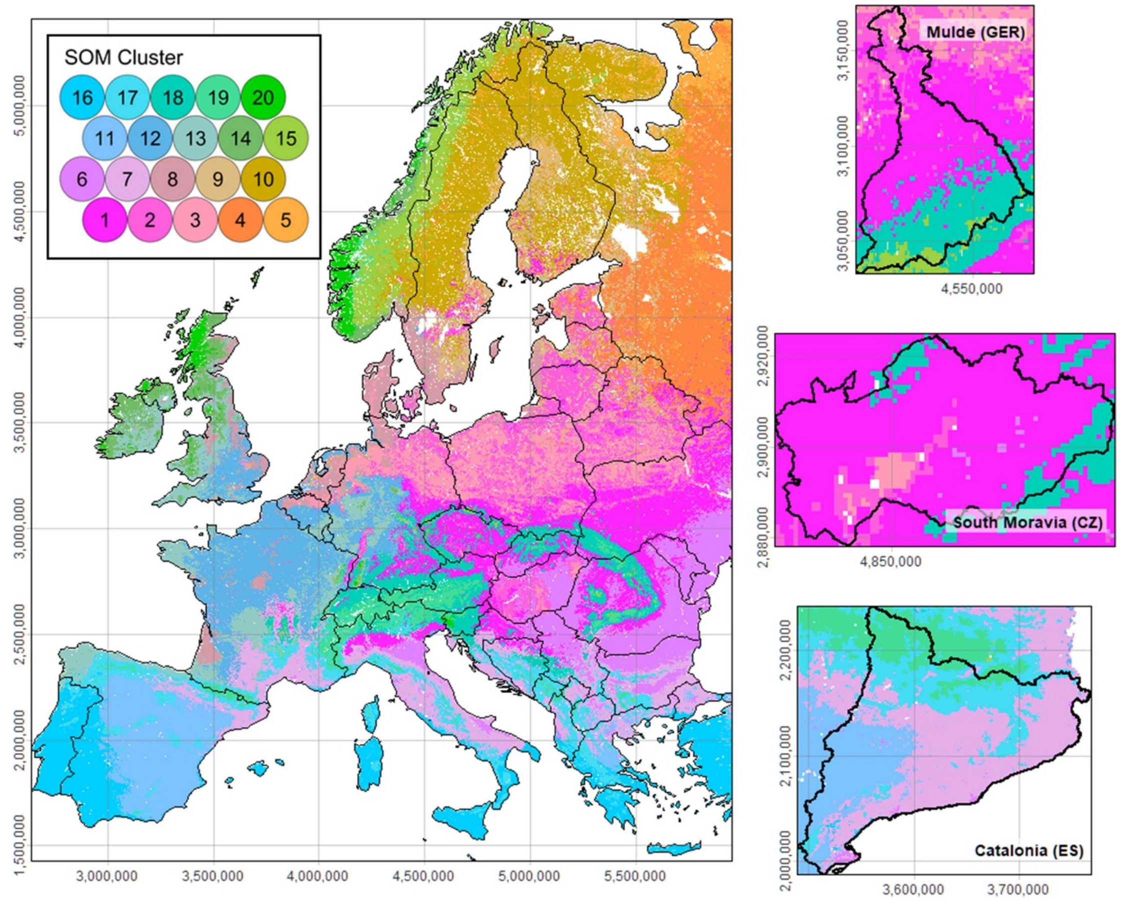

Figure 3. SOM *k*20 cluster map of the study area with color-coded clusters. Right side SOM *k*20 clusters in the regional case studies. At this scale, the case study regions were divided into five, six, and five clusters for Czechia, Spain and Germany, respectively.

2020) that is part of the Integrated Administration and Control System. We selected 'arable land' field parcels, excluding parcels with permanent grassland. The Czech field information was extracted from the public Land Parcel Identification System (LPIS) of the Ministry of Agriculture of the Czech Republic, combining the categories 'arable land' and 'grassland on arable land'. For Spain, the LPIS data provided by the Centre for Ecological Research and Forestry Applications was restricted to the 'arable land' category. All three datasets were rasterized, first to a 10 m spatial resolution (to preserve finer detail) and subsequently aggregated to a 100 m resolution. Concurrently, the SOM had to be disaggregated from 1 km to 100 m resolution using the *disaggregate* function from the R-package *raster*.

3. Results

3.1. Continental application—SOM *k*20

The identified archetypes of agri-environmental potential showed a relatively even geographical distribution and their coverage ranged from 1.0% (Cluster 20 with 62 000 km^2) to 10.1% (Cluster 10 with 640 000 km^2) of European land (figure 3). The largest clusters, 4 (542 000 km^2) and 10 (640 000 km^2), were in Northern Finland and Russia, suggesting that there is a relatively homogenous space of environmental conditions over a large area, although much of it with low agricultural potential. The highest QE was found in clusters 19 and 20 (figure 4), located along the coast of Norway and the northern UK, and also at the coast of Spain, Portugal and the Alpine region. These archetypes were the most heterogeneous, clustering agri-environmental potential with a wide range of conditions, especially elevation and precipitation (figure A7).

An important output of the SOM procedure is so-called heatmaps (or component planes), which are depictions of the relative contribution of each input variable to the overall ordering of the SOM output space (figure 4). Comparing multiple heatmaps reveals non-linear and partial correlations between variables, providing a cross-sectional view of our 17 input variables. For example, elevation and terrain ruggedness showed a similar pattern of high values towards the top of the plane (especially cluster 19), descending towards low values in the bottom part of the plane. Conversely, archetypes associated with high values of soil bulk density or clay content at the

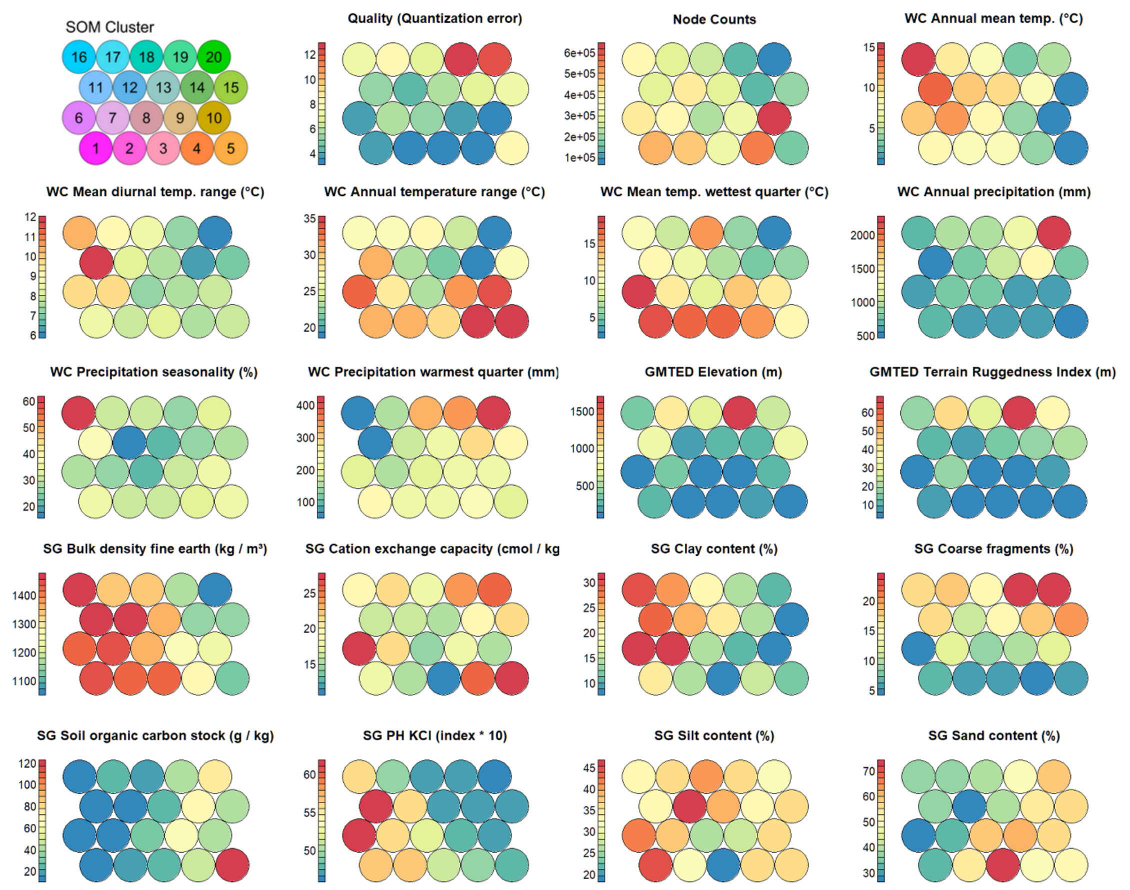

Figure 4. Heatmaps of input variable distribution across the SOM *k*20 grid, showing the relative contribution of each input variable to the overall ordering of the self-organizing map.

left part of the plane and decreasing values to the right showed the opposite pattern in terms of soil organic carbon or sand content.

The comparison of the identified archetypes with independent agricultural data (IIASA field sizes and mean cropland cover) showed that even coarse-scale clustering may have a meaningful agricultural relevance (figure 5). For example, the ordering of identified agri-environmental archetypes captured a pattern of decreasing field sizes going from the bottom to the top portion of the SOM grid and decreasing cropland cover going from left to right. All categories of field sizes tended to occur in archetypes with higher cropland cover but archetypes with a high proportion of no fields only partly coincided with low cropland cover, likely because the global field size data also included permanent pastures.

3.2. Regional application—SOM *k*400

Unsurprisingly, the regional application clustered European land into 400 smaller and more homogeneous agri-environmental archetypes than in the case of SOM *k*20 (figure 6). The sizes of clusters ranged from 2230 km^2 (0.04% of the study area) for cluster 381–34 000 km^2 (0.5% of the study area) for cluster 184, with a median of 15 068 km^2, which is close to 1/400 of the total study area. Smaller clusters tended to be less heterogeneous (lower QE), but the overall cluster quality was uniformly distributed across Europe and higher than in the case of *k*20 (figures 7 and 8). A correlation of input variables with the clusters' mean QE (figure A9) showed that QE was positively associated with annual precipitation, soil coarse fragments, terrain ruggedness and elevation. Therefore, agri-environmental potential with high values of these variables, located along the coast of Norway, Northern UK and the Alpine region, were also more heterogeneous and thus less likely to form homogeneous archetypes.

SOM heatmaps exhibited much more distinctive patterns and many fewer correlations between input variables than in the *k*20 case, suggesting the *k*400 clustering provided a more detailed typology of agri-environmental systems. However, some patterns were consistent as in the continental application. For example, elevation, terrain ruggedness and precipitation show a pattern of high values towards the top of the plain, while several soil characteristics, such as bulk density, clay content, or soil organic carbon show a left-to-right distribution of values.

The SOM *k*400 clustering was also able to better capture the spatial pattern in the independent agricultural data than in the *k*20 case (figure 8). The field sizes tended to increase when going from the bottom

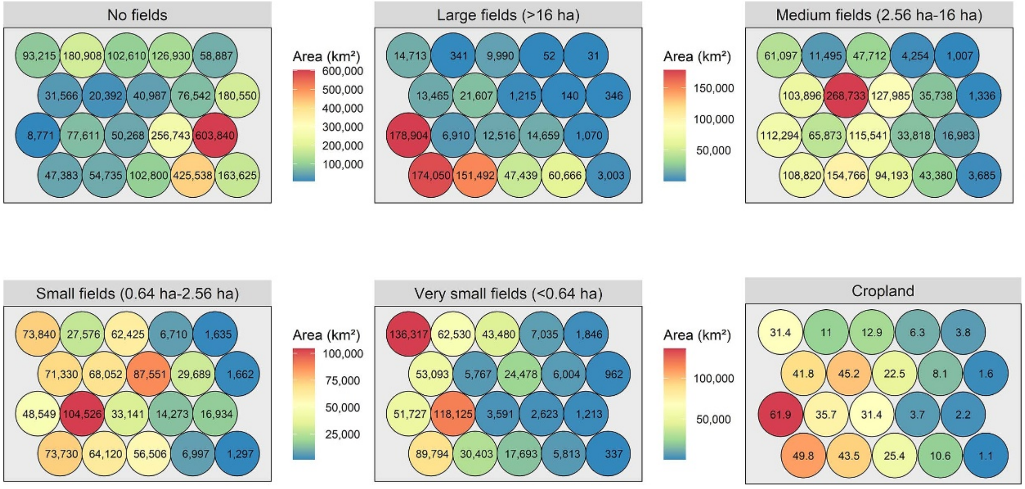

Figure 5. SOM *k*20 node grid with cluster distributions of IIASA field size areas and mean cropland cover.

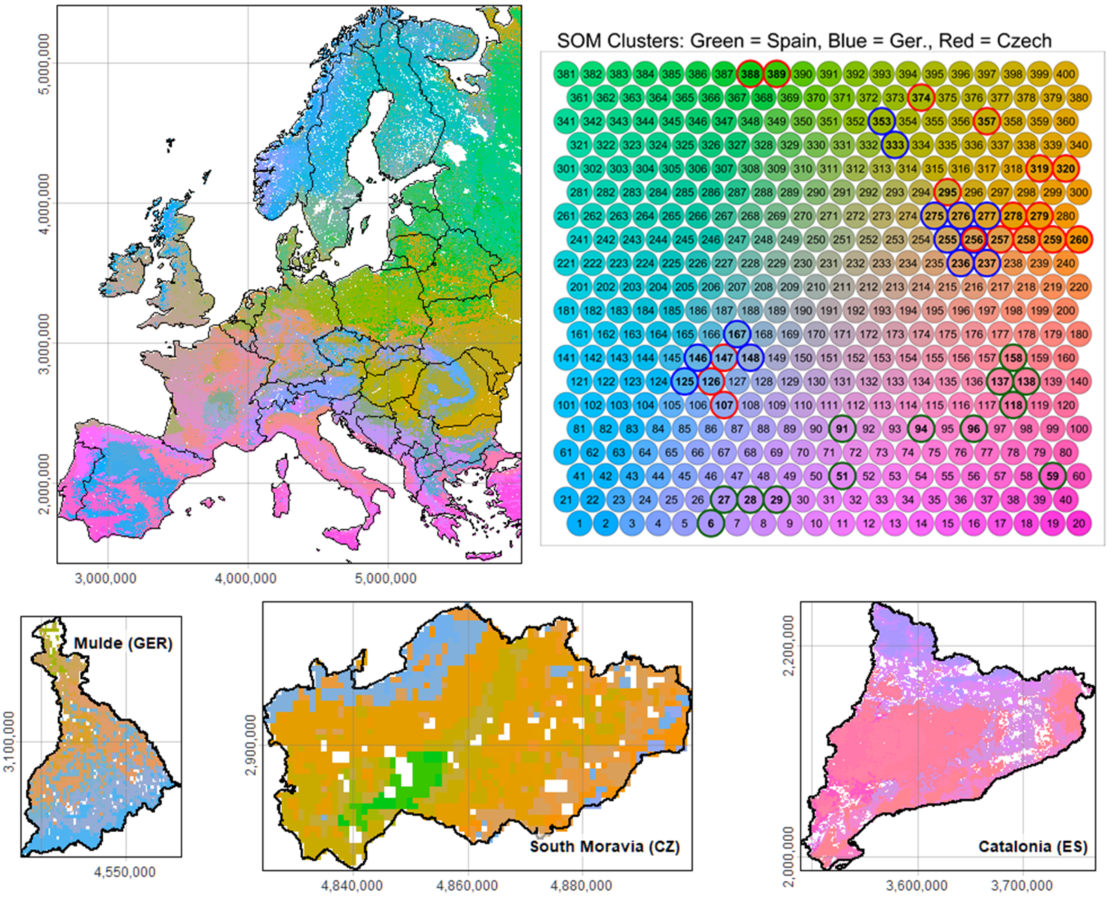

Figure 6. SOM *k*400 node grid with color-coded cluster numbers. Nodes with red outlines correspond to the clusters found in the Czech case study area, green in Spain and blue in Germany while disregarding clusters that covered less than 1% of the total case study area. The color code for the actual map (left) can be found on the right side. It was color-coded in a four-way gradual color scheme, to emphasize that clusters close to each other share similar characteristics, due to the topological nature of SOMs.

to the top in the SOM grid, following similar pattern as in several input variables, e.g. elevation, terrain ruggedness and soil coarse fragments. The regional-scale application also captured a clearer pattern in the cropland cover distribution in Europe, with agri-environmental archetypes identified on the left of the SOM grid having low cropland cover, but those on the right having a high cropland cover.

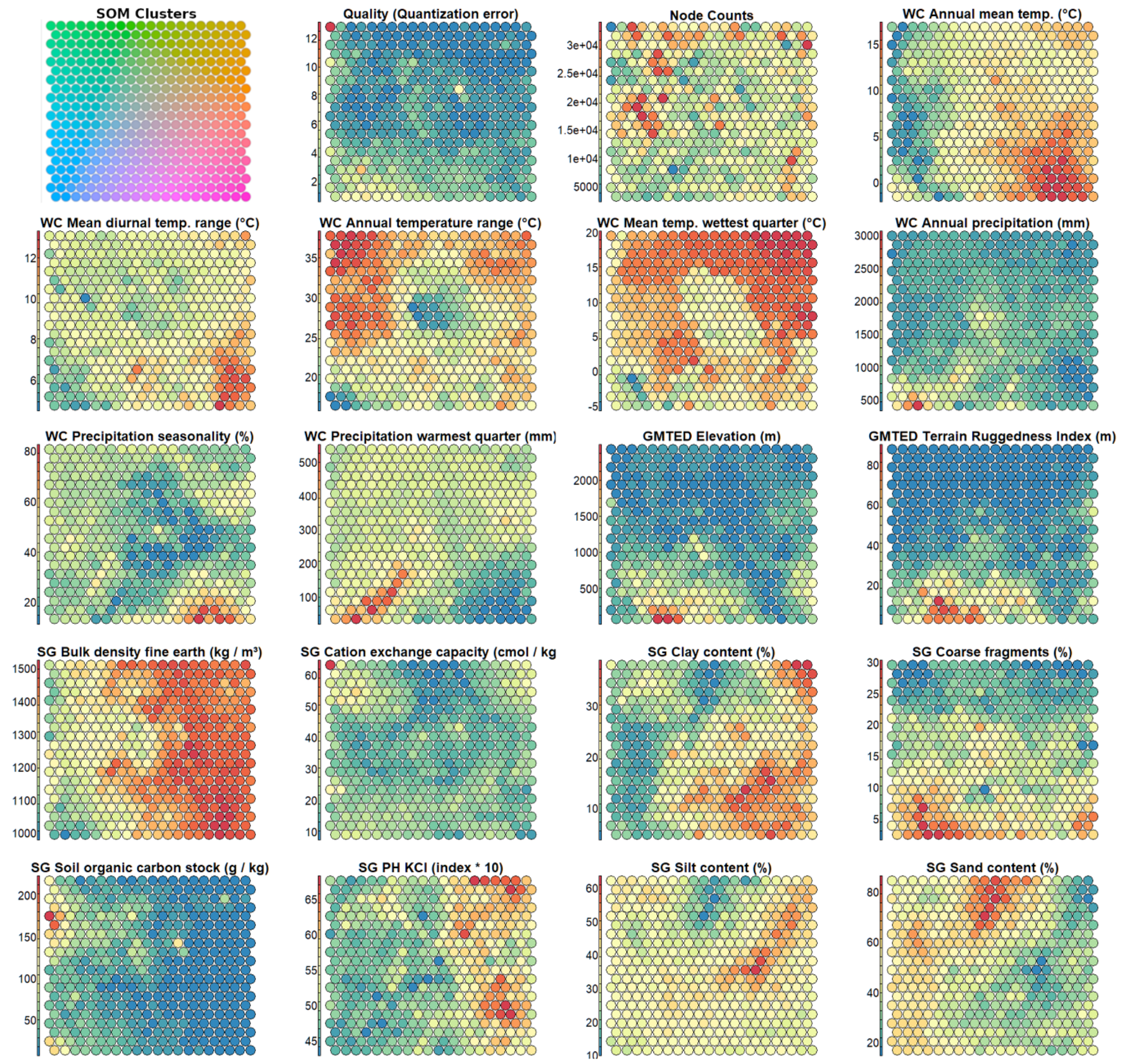

Figure 7. Heatmaps of input variable distribution across the SOM *k*400 grid.

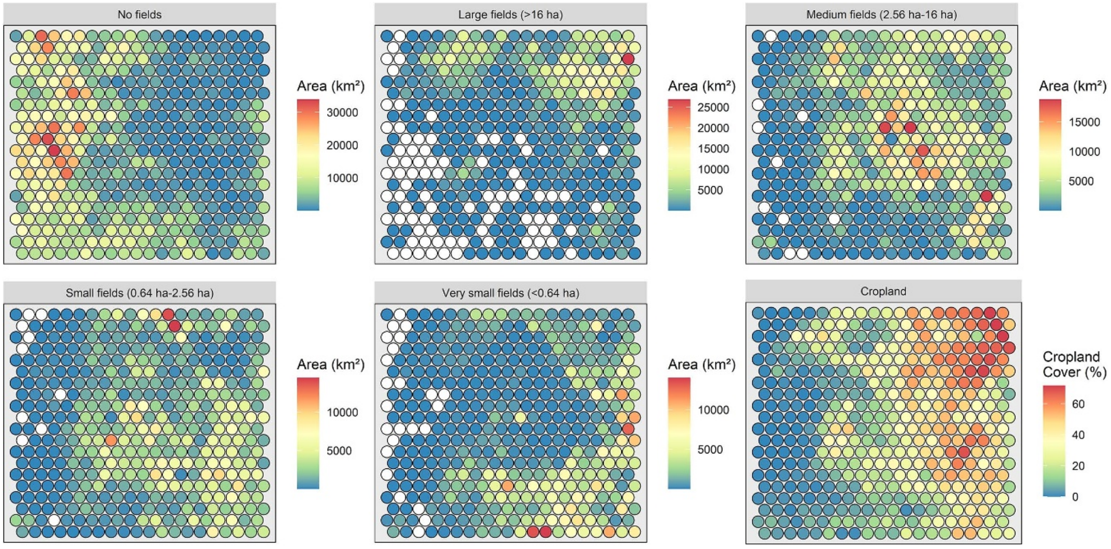

Figure 8. SOM *k*400 node grid with cluster distributions of IIASA field size areas and mean cropland cover.

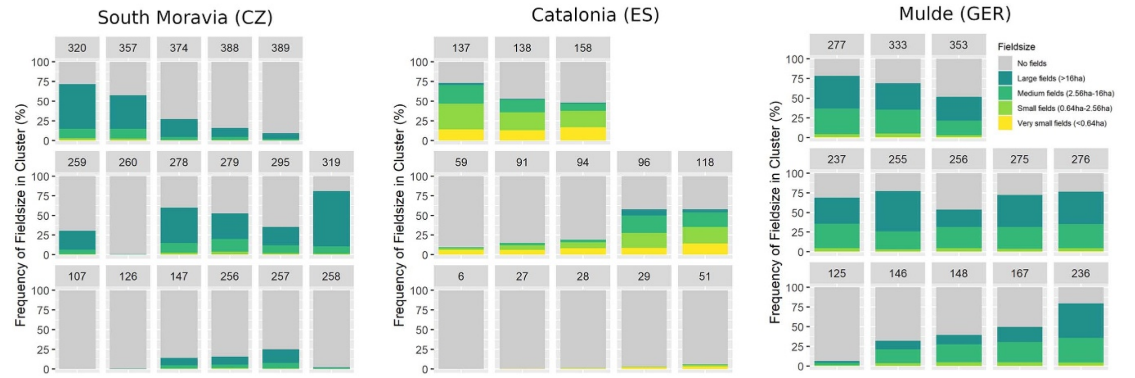

Figure 9. Frequency of field sizes per SOM *k*400 clusters in the three regional case studies. Only clusters larger than 1% of the total case study area are shown.

3.3. Regional case studies

The regional-scale clustering was better suited to identifying detailed agri-environment archetypes in the case study regions. While the *k*20 classification identified 4–5 archetypes in each region, typically capturing the main climate and elevation gradients, the *k*400 classification identified 13–17 archetypes in each region (disregarding the few clusters that covered less than 1% of the total case study area) (figure 6). Because of the small area of the Czech region, a large fraction of clusters shared relatively similar environmental characteristics. However, clusters with different proportions of large- and medium-size fields versus no field coverage were still well distinguished (figure 9). Similarly, the 13 distinct clusters in the Spanish case study showed a clear differentiation of cropland frequencies. In contrast, the German case study had the majority of land used as cropland and the clusters showed a relatively equal distribution of large and medium fields within clusters. The only exceptions were the archetypes in the very south of the case study that had a lower proportion of cropland and a higher proportion of permanent grassland. The archetypes in the German case study were driven largely by the strong north-south gradient of elevation, climate and soil conditions that did not coincide with field size distribution.

4. Discussion

This study provides an illustrative, data-driven approach for identifying and mapping archetypes of agri-environmental potential in Europe. Our work extends previous efforts creating agri-environmental typologies in that it (a) considers recent, agriculturally-important, biophysical variables that have not been previously available at the European extent and (b) is based on a fully data-driven, unsupervised clustering approach that eliminates potential biases typically associated with expert-driven or supervised techniques used to define classification thresholds. By applying this method to 17 key indicators at two spatial scales (*k*20 and *k*400), we demonstrate the scalability of our approach to generalize the complexity of environmental conditions relevant for agriculture at European and regional scales, respectively. We gained insight into the agricultural relevance of identified archetypes by comparing them with independent data on cropland cover and field size across Europe but especially in three regional case studies in Germany, Czechia and Spain.

The spatial classification of agri-environmental archetypes presented here includes four main biophysical and climatic determinants of agricultural production capacity; precipitation, temperature, topography and soil characteristics. Climate and weather exert significant influence on agricultural production and by extension decisions on land use and distribution of agricultural activities. In Europe, approximately 60%–70% of annual yield variation for major crops (i.e. wheat, sugar beets) can be attributed to weather conditions (Trnka *et al* 2016). Soil properties have driven the decisions where different types of agriculture are implemented and where land-use change occurs, both historically (e.g. Ellenberg 1990) and until the present day (e.g. van Vliet *et al* 2015, Meyer and Früh-Müller 2020). At the same time, soils are also heavily influenced by climatic factors, geology and agricultural activities (Hengl *et al* 2017). Therefore, the omission of soil information from the classification of agricultural typologies could be seen as a potential shortcoming of previous classification approaches. This may limit their suitability for supporting decision-making within the context of agriculturally used lands, although direct comparison with our classification would be needed to determine how substantial the difference is. Our approach sought to overcome this limitation by including recent, high-resolution soil information (SoilGrids). Our analyses show that using these input data allows for a good differentiation of cropland cover and field size both at continental and regional scales. Nevertheless, using these environmental variables alone will not be sufficient since variables like field size and cropland cover are also influenced by socio-economic, historic and political factors (e.g. Batáry *et al* 2017, Sroka *et al* 2019). The spatial classification of agri-environmental potential presented here seeks to represent the fundamental, environmental background within which any other land-use decision is embedded and within which societal aspects influence land-use decisions.

The presented typology was developed with improving the spatial targeting of agricultural policy and agri-environmental management in mind. Agricultural policies, such as those derived from the EU's Common Agricultural Policy, tend to ignore the complexity of Europe's agricultural systems, leading to inconsistent and uncertain outcomes in different locations (Ziv *et al* 2020). Addressing territorial diversity by mapping archetypes with similar agri-environmental conditions is a crucial step towards tailoring policies that would fit national or regional needs. For instance, our approach can assist in deciding where specific agri-environmental schemes or practices (e.g. no tillage) may be appropriate and should therefore be subsidized, given the agri-environmental potential in the area. However, we envision our approach to be useful in many other applications (figure 1). For example, our approach can be used to stratify regions for selection of research sites or to assess geographical representativeness and spatial bias in existing research site networks (Wohner *et al* 2021). Agri-environmental archetypes can also serve as a modeling framework for investigating the interactions between farming and biodiversity in different types of agricultural systems (Seppelt *et al* 2020, Jungandreas *et al* 2022), for assessing bundles of ecosystem services (Cord *et al* 2017) or for modeling the complexity of decision making and behavior of different agents in the agricultural sector (Will *et al* 2021). Thanks to its scalable character, the approach is especially suited for upscaling of land-use models and management recommendations developed in regional case studies to larger geographical extents and for assessing the transferability of place-based research to other regions with similar agri-environmental characteristics (Václavík *et al* 2016).

More broadly, our study contributes to the burgeoning field of archetype analysis in sustainability research (Oberlack *et al* 2019, Eisenack *et al* 2021). We used a machine-learning clustering (i.e. SOMs), rated among the most promising techniques in the methodological portfolio of archetype analysis (Sietz *et al* 2019), allowing for the comparison of typical variable combinations both in terms of similarity and, when applied to spatial data, geographic proximity. Such approach allows synthesizing general patterns of land systems, and consequently building middle range theories that stand between simplistic descriptions of singular cases (e.g. case studies or grid cells as in our study) and universal theories, providing a pathway towards a more generalized knowledge in land system science (Meyfroidt *et al* 2018, Rocha *et al* 2020). This typology of cases also enhances the treatment of causality in archetype analysis (Sietz *et al* 2019), going towards 'thick description' (more quantitative insights into recurrent features) and 'causal factor configurations' (insights into patterns of archetype determinants), as it is using high-dimensional data and is applicable at multiple spatial or temporal scales.

Besides being a typology of cases where each case (land area, grid cell, etc) is classified as exactly one archetype, archetypes can be also seen as building blocks of dynamic systems, representing causal mechanisms that explain individual cases (Oberlack *et al* 2019). A combination of both complementary approaches has been recommended as a fruitful avenue to follow (Eisenack *et al* 2021). For example, our typology can be used as a starting point to identify regions with similar agri-environmental potential suitable for a certain policy, but government efforts to implement the policy may be effective in certain socioeconomic contexts but less effective or even counterproductive in others. Therefore, using data on farmer characteristics, stakeholder demands or economic background may allow identifying archetypal causal mechanisms between policies and agricultural sustainability, which may ultimately help more effectively transfer policies across geographical and social contexts.

Our results are limited by methodological requirements of our approach. We assessed the quality of our classification procedure by calculating the QE (i.e. the distance of each grid cell in the multidimensional space of variables to the mean variable values that characterize each archetype). It shows a pattern of relatively short distances for most locations, indicating robust typology figures A7 and A8). However, due to the methodological requirement to draw the first initialization of weight vectors randomly, the outputs of different SOM runs are never fully identical. This is a known, potential problem in the analysis of complex, high-dimensional data (Mariette and Villa-Vialaneix 2016). A possible solution in the case of variable results between runs is combining multiple runs while preserving the topological properties of SOM, e.g. by bootstrapping or hierarchical clustering techniques (Petrakieva and Fyfe 2003, Mariette and Villa-Vialaneix 2016). Another option is to obtain the initial weights and position of prototype-nodes from the first two eigenvectors of a principal component analysis (PCA) performed on the matrix of input variables (Ciampi and Lechevallier 2000), but the abstract axes of PCA hinder consequent interpretation of clustering results. While these possible variations between individual runs do not impact the applicability of the resulting outputs, they should be carefully considered when developing typologies based on SOM.

Our typology is also limited by the selection of input variables. In total, we used 17 variables that cover the range of biophysical variation of agricultural systems in Europe but all have different levels of uncertainties associated with them, depending on the origin and scale of the data. While alternative databases exist with potentially higher resolution, e.g. EU-digital elevation model (DEM) for elevation or land use/cover area frame statistical survey (LUCAS) Topsoil for soil properties, they are limited in their geographic coverage or in their consistency across European countries. Additionally, while extreme events like droughts and frosts are highly relevant for agricultural systems, no Europe-wide datasets are available on the probability of extreme conditions at resolutions below 0.25 degrees (e.g. E-OBS dataset from the Copernicus Climate Change Services). Potentially, this lack of data can be alleviated by using phenology indices that capture variation in the vegetation period. Indeed, we created the GDD variable as a proxy for the vegetation period but it was not selected due to its strong correlation ($r = 0.98$) with the mean annual temperature. However, such correlation may not be present at different spatial scales or for more specific phenological indices produced from direct phenological observations (Chmielewski and Rötzer 2001). Therefore, spatially-explicit data that may arise from a harmonized phenological survey across all of Europe (COST action 725), which led to the Pan European Phenology Database (PEP725l; Templ *et al* 2018), have a high potential for future improvements of agri-environmental typologies.

5. Conclusion

This study has identified spatially-explicit archetypes of agri-environmental potential in Europe, focusing on two spatial scales: a continental scale and a regional scale. The two typologies captured the main biophysical variation of agriculture systems in Europe thanks to the SOMs' capability to visualize multidimensional data, thus fostering the interpretation of their agricultural relevance. However, our approach can be adapted and scaled to fit the requirements of other scales and study contexts. In addition to serving as a spatial framework for tailoring agricultural policies and management, we see the main application domains in site selection and stratification, modeling of interactions between agriculture and ecosystem services, and in assessing the transferability of agriculture-relevant models to other regions and across scales. The recently started war in Ukraine has further highlighted the need for Europe's agricultural sector to be able to respond quickly and in a spatially targeted manner to mitigate crises and ensure food security. Future efforts could also recreate agri-environmental archetypes with climate scenarios instead of historical climate data. Predicting the potential spatial change of agri-environmental patterns can help anticipate how agricultural policies may need to be adapted in the future due to climate change.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://geonetwork.ufz.de/geonetwork/srv/eng/catalog.search#/metadata/3e2df2bd-b98e-4854-88a2-d7555a36cc22. Data will be available from 1 October 2022.

Acknowledgments

This work was supported by the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 817501 (BESTMAP; Ziv *et al* 2020). We thank Alexander Brenning for helpful comments on the study design.

Agri-environmental archetypes at the continental scale (SOM k20) are available at: https://geonetwork.ufz.de/geonetwork/srv/eng/catalog.search#/metadata/3e2df2bd-b98e-4854-88a2-d7555a36cc22. Agri-environmental archetypes at the regional scale (SOM k400) are available at: https://geonetwork.ufz.de/geonetwork/srv/eng/catalog.search#/metadata/fb6027c3-2a51-4c5e-aa19-a3ef86400490. The computer code to run the SOM classification is available at: https://git.ufz.de/bestmap/agri-environmental-potential.

Appendix

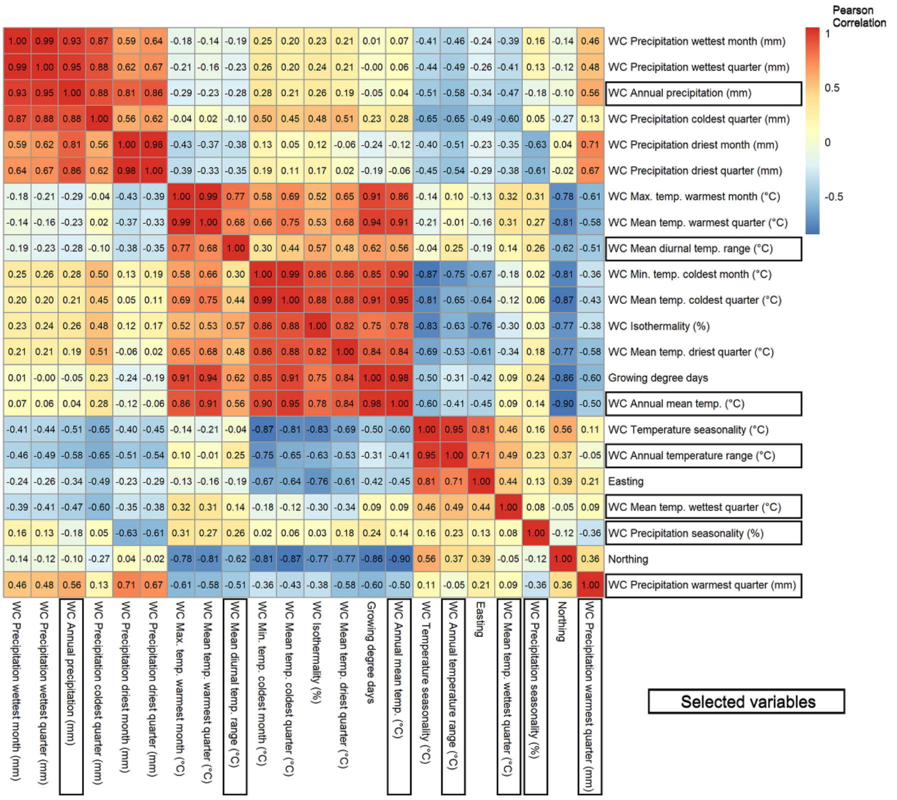

Figure A1. Pearson correlation matrix for the climate and location variables. The selected variables were used in the clustering process.

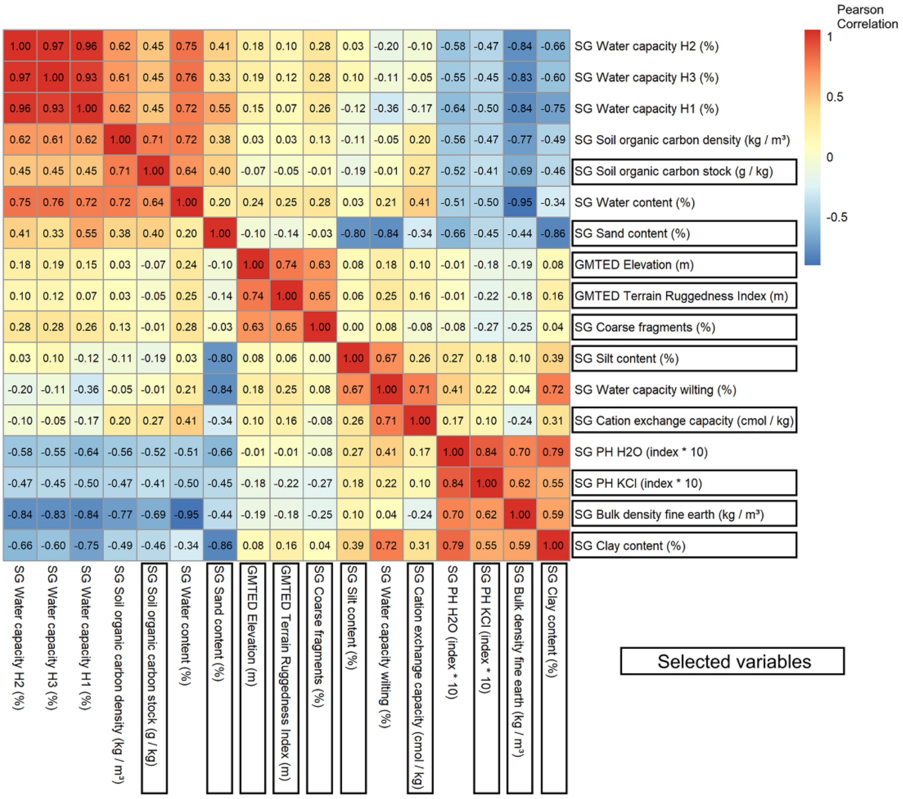

Figure A2. Pearson correlation matrix for the soil and topographic variables. The selected variables were used in the clustering process.

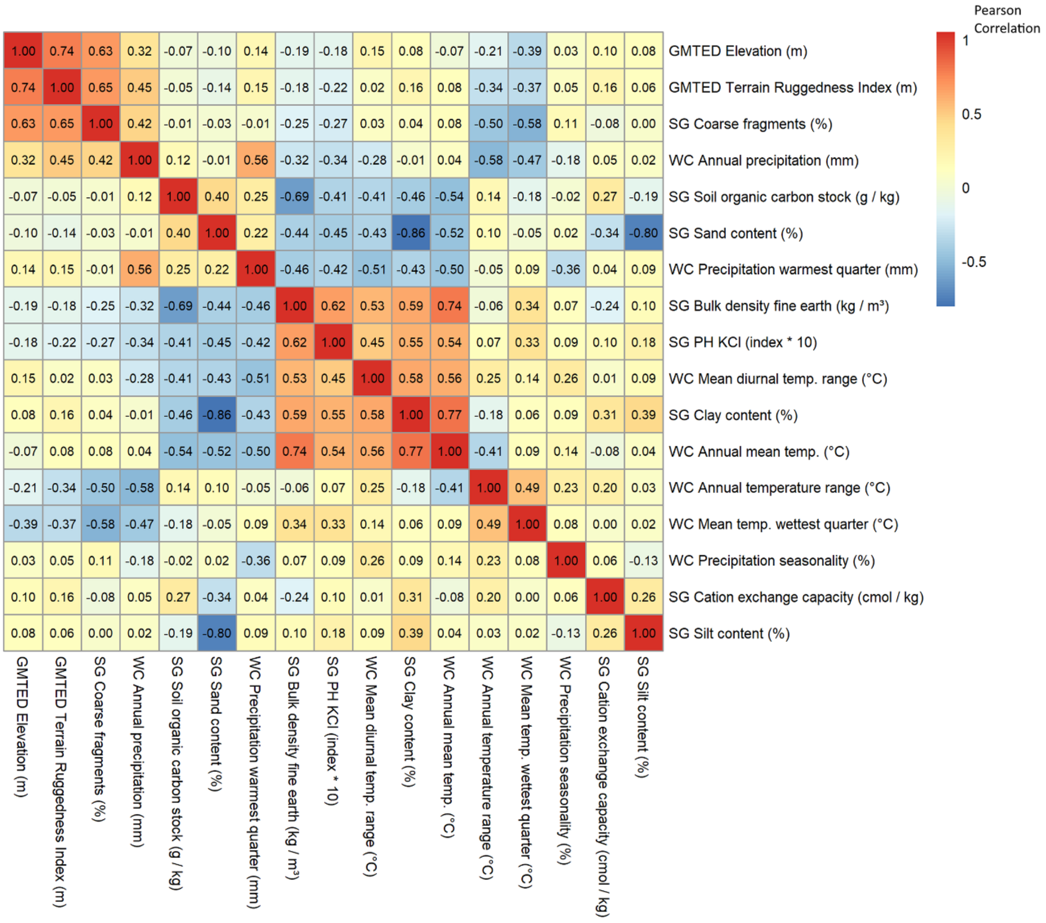

Figure A3. Pearson correlation matrix for final input variables used in the clustering process.

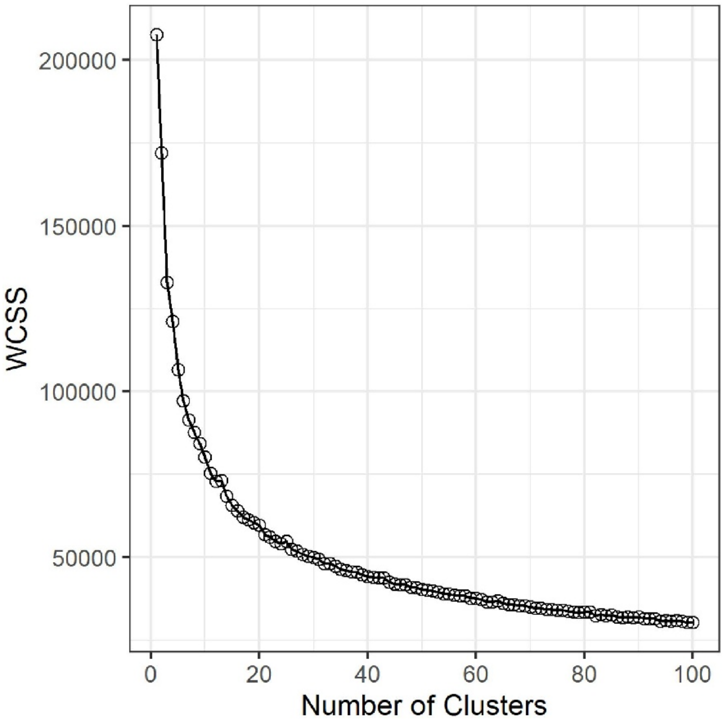

Figure A4. Within-cluster sum of squares (WCSS) for differently sized HAC clusters of the SOM outputs.

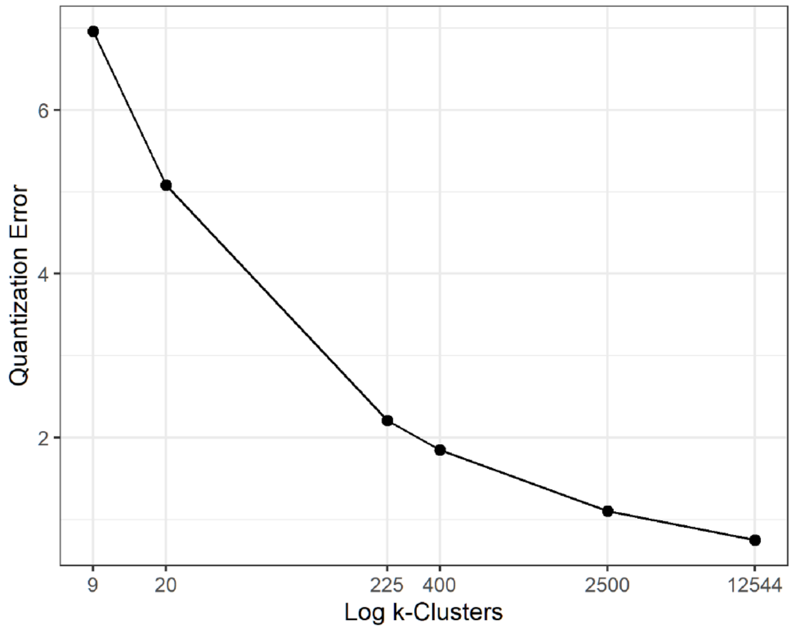

Figure A5. Mean quantization errors for multiple runs of differently sized SOM clustering results.

Table A1. Pearson's correlation coefficients between each input variable and the agricultural data (cropland cover and field size, respectively). The calculation is based on 1% of randomly selected pixels to avoid spatial autocorrelation.

Variable name (unit)	Cropland cover	Field size
Annual precipitation (mm)	−0.11	0.09
Precipitation warmest quarter (mm)	−0.22	−0.19
Mean Diurnal temp. range (°C)	0.35	0.34
Annual mean temp (°C)	0.54	0.61
Annual temp. range (°C)	−0.12	−0.31
Mean temp. wettest quarter (°C)	0.31	0.14
Precipitation seasonality (%)	−0.08	−0.09
Coarse Fragments (%)	−0.25	−0.03
SOC concentration (g kg^{-1})	−0.55	−0.49
Sand content (%)	−0.47	−0.41
Bulk density (kg m^{-3})	0.62	0.56
PH KCl (index *10)	0.67	0.43
Clay content (%)	0.48	0.49
Cation exchange capacity (cmol kg^{-1})	−0.05	−0.11
Silt content (%)	0.26	0.13
Elevation (m)	−0.12	−0.02
Terrain Ruggedness (m)	−0.17	0.05

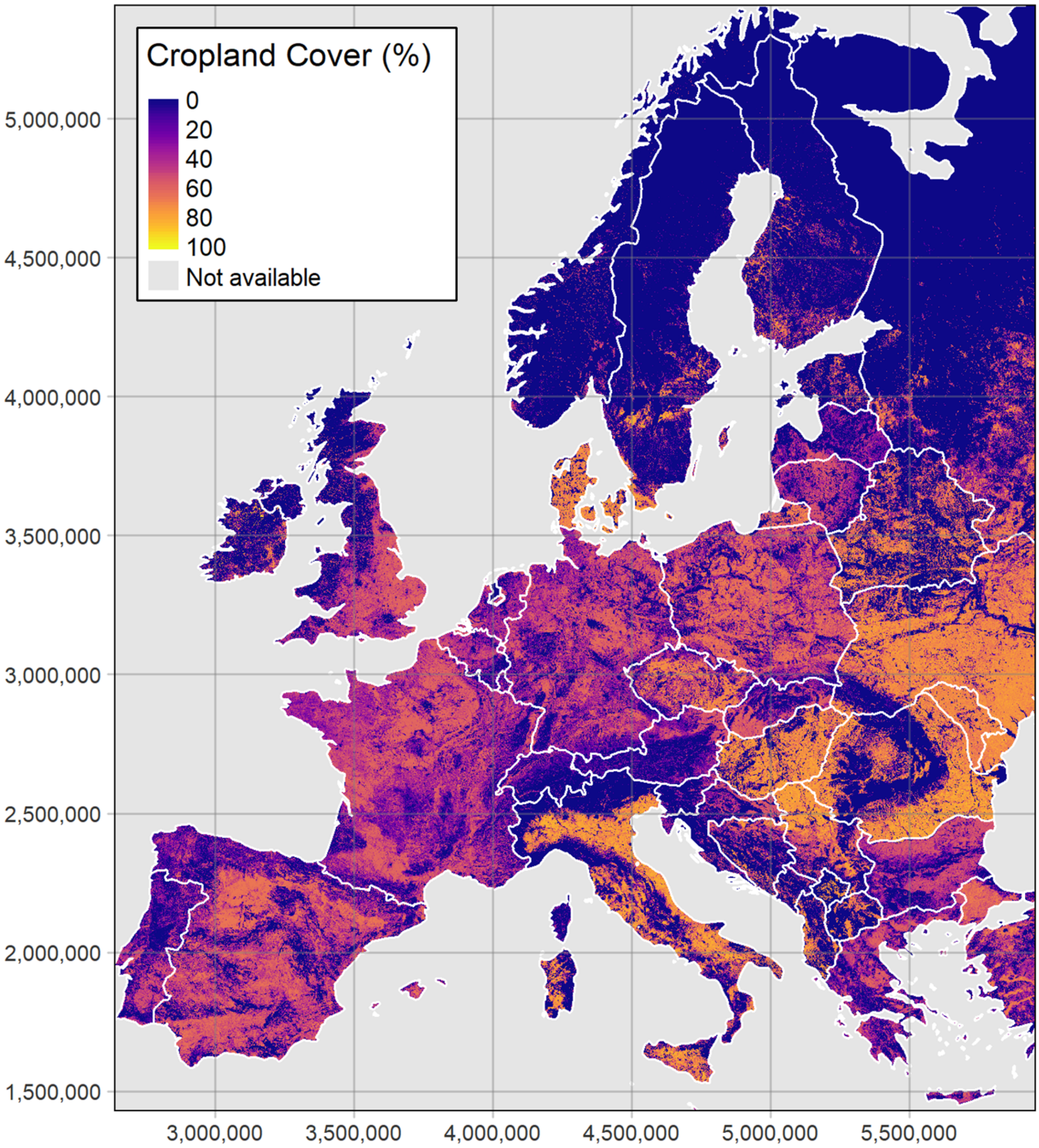

Figure A6. IIASA-IFPRI cropland percentage map with a 1 km resolution.

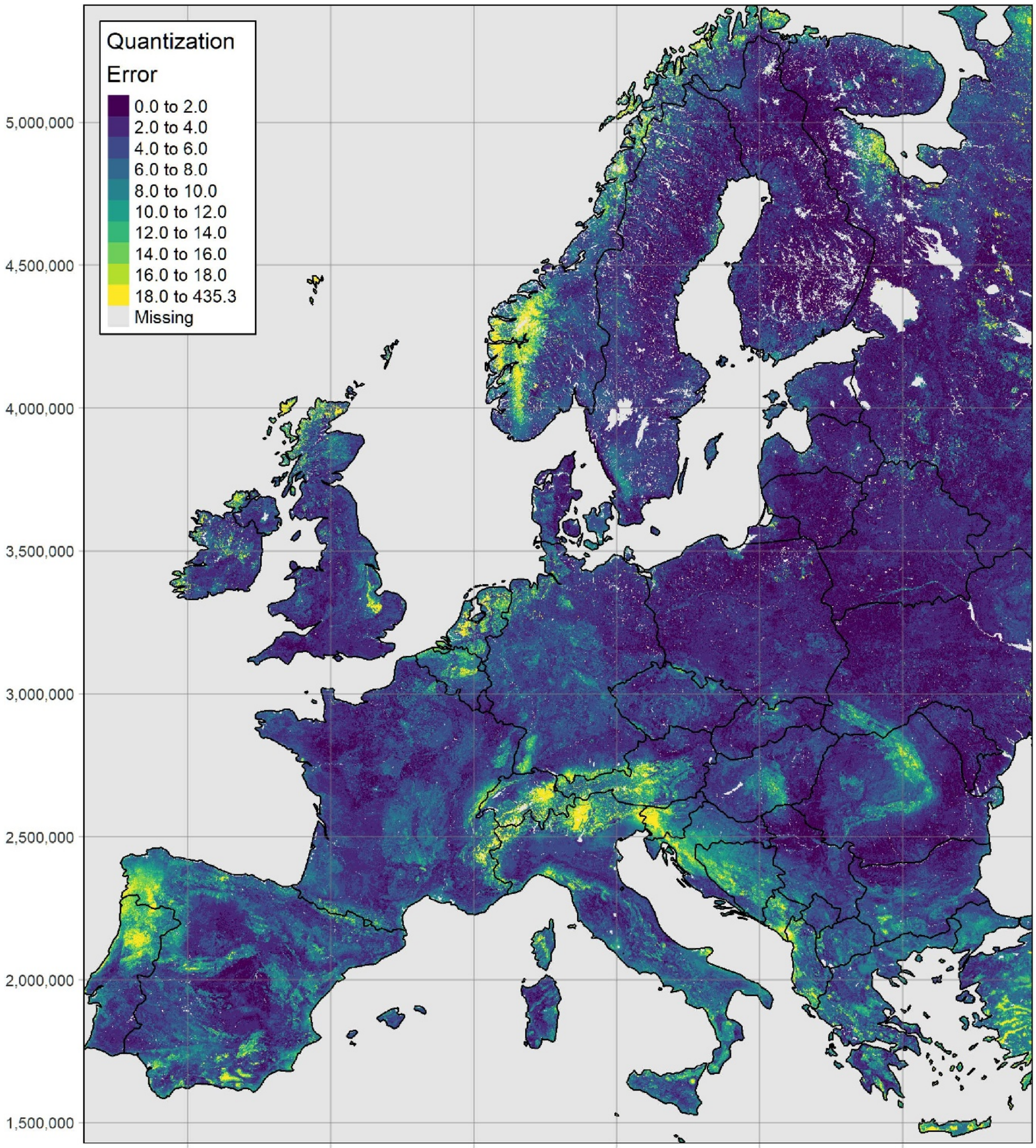

Figure A7. The SOM *k*20 quality map of quantization error (QE), i.e. the distance for each grid cell to its corresponding cluster centroid.

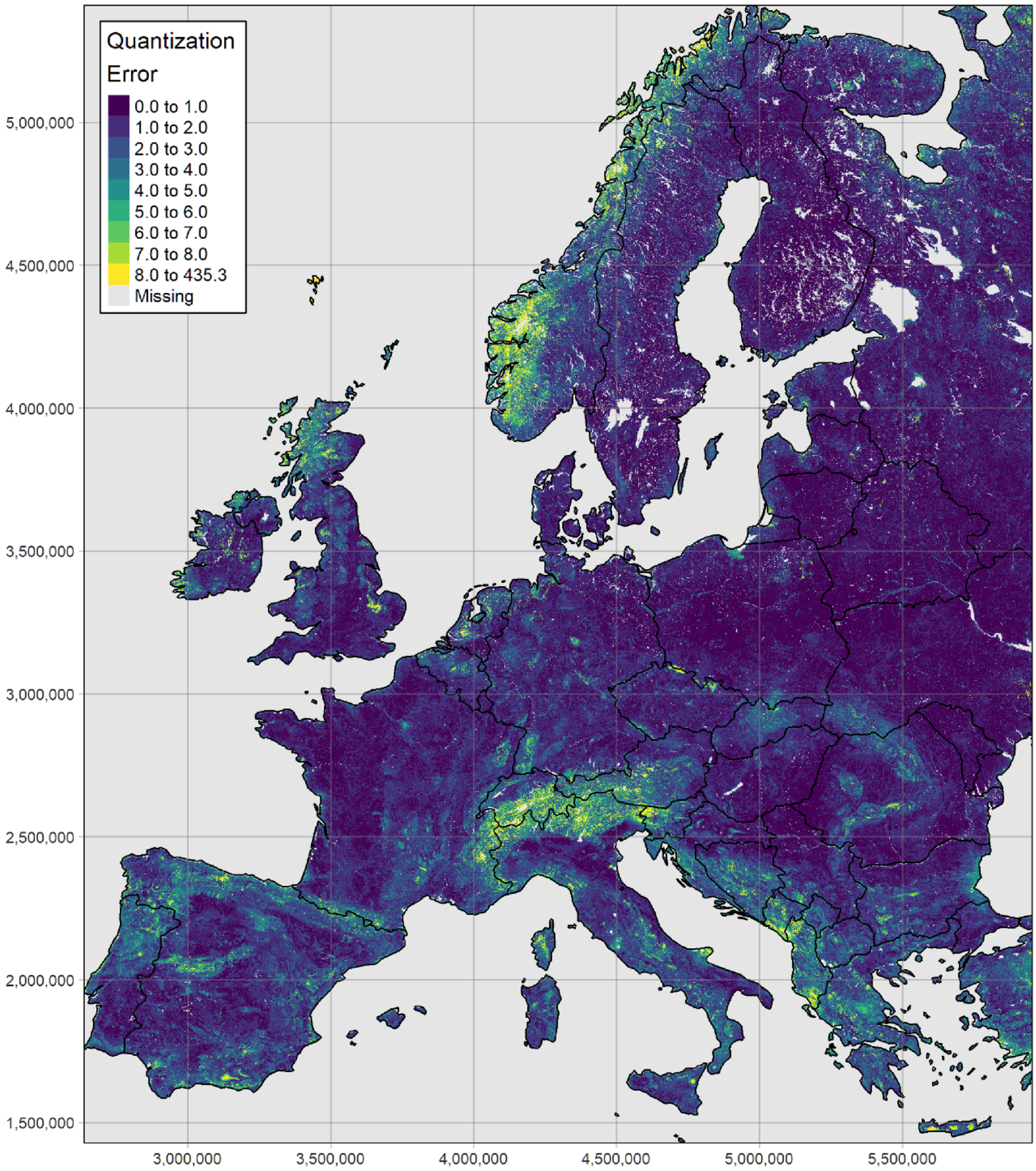

Figure A8. The SOM *k*400 quality map of quantization error (QE), i.e. the distance for each grid cell to its corresponding cluster centroid.

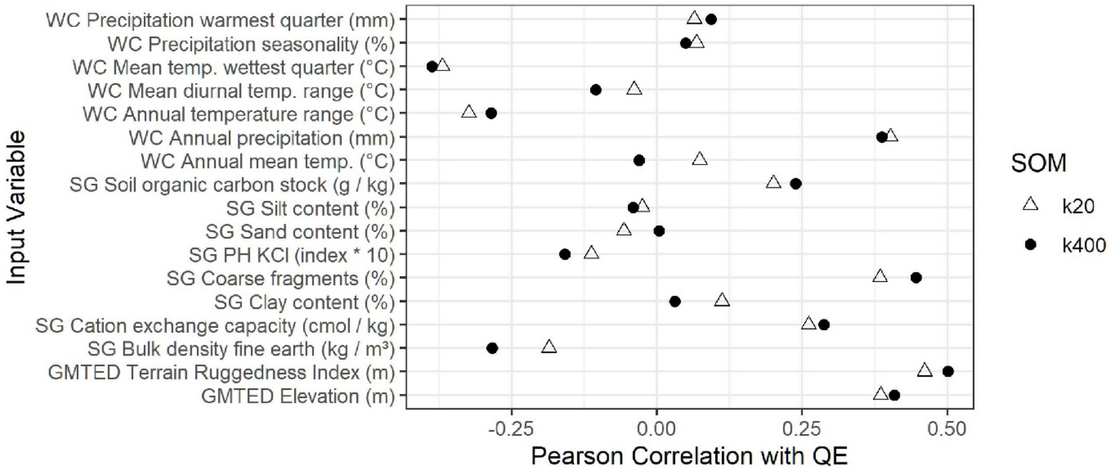

Figure A9. Pearson correlations of all data points in the input variables with the quantization errors for both cluster sizes (*k*20 and *k*400).

ORCID iDs

Michael Beckmann https://orcid.org/0000-0002-5678-265X
James M Bullock https://orcid.org/0000-0003-0529-4020
Anna F Cord https://orcid.org/0000-0003-3183-8482
Guy Ziv https://orcid.org/0000-0002-6776-0763
Tomáš Václavík https://orcid.org/0000-0002-1113-6320

References

Amatulli G, Domisch S, Tuanmu M-N, Parmentier B, Ranipeta A, Malczyk J and Jetz W 2018 A suite of global, cross-scale topographic variables for environmental and biodiversity modeling *Sci. Data* **5** 180040

Andersen E 2017 The farming system component of European agricultural landscapes *Eur. J. Agron.* **82** 282–91

Batáry P *et al* 2017 The former Iron Curtain still drives biodiversity-profit trade-offs in German agriculture *Nat. Ecol. Evol.* **1** 1279–84

Batáry P, Dicks L V, Kleijn D and Sutherland W J 2015 The role of agri-environment schemes in conservation and environmental management *Conserv. Biol.* **29** 1006–16

Bureau J-C, Tangermann S, Matthews A, Viaggi D, Crombez C, Knops L and Swinnen J 2012 The common agricultural policy after 2013 *Interecon. Econ. Rev. Eur. Econ. Policy* **47** 316–42

Castillo C P, Kavalov B, Diogo V, Jacobs-Crisioni C, Silva F B E, Baranzelli C and Lavalle C 2018 Trends in the EU agricultural land within 2015–2030 *JRC Working Papers* (No. JRC113717) (Joint Research Centre) (Seville site)

Chmielewski F-M and Rötzer T 2001 Response of tree phenology to climate change across Europe *Agric. For. Meteorol.* **108** 101–12

Ciampi A and Lechevallier Y 2000 Clustering large, multi-level data sets: an approach based on Kohonen self organizing maps *Proc. 4th European Conf. on Principles of Data Mining and Knowledge Discovery, PKDD 2000* (Berlin: Springer-Verlag) pp 353–8

Cord A F *et al* 2017 Towards systematic analyses of ecosystem service trade-offs and synergies: main concepts, methods and the road ahead *Ecosyst. Serv.* **28** 264–72

Cracknell M J, Reading A M and de Caritat P 2015 Geological knowledge discovery and minerals targeting from regolith using a machine learning approach *ASEG Ext. Abstr.* **1** 1–4

Delmelle E, Thill J-C, Furuseth O and Ludden T 2013 Trajectories of multidimensional neighbourhood quality of life change *Urban Stud.* **50** 923–41

Dittrich A, Seppelt R, Václavík T and Cord A F 2019 Spatial patterns of ecosystem service bundles in Germany *Atlas of Ecosystem Services: Drivers, Risks, and Societal Responses* ed M Schröter, A Bonn, S Klotz, R Seppelt and C Baessler (Cham: Springer International Publishing) pp 279–83

Dormann C F *et al* 2013 Collinearity: a review of methods to deal with it and a simulation study evaluating their performance *Ecography* **36** 27–46

Eisenack K, Oberlack C and Sietz D 2021 Avenues of archetype analysis: roots, achievements, and next steps in sustainability research *Ecol. Soc.* **26** 31

Ellenberg H 1990 *Bauernhaus und Landschaft in ökologischer und historischer Sicht* (Stuttgart: Ulmer E)

Fick S E and Hijmans R J 2017 WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas *Int. J. Climatol.* **37** 4302–15

Fritz S *et al* 2015 Mapping global cropland and field size *Glob. Change Biol.* **21** 1980–92

Galdies C and Vella K 2019 Future climate change impacts on Malta's agriculture, based on multi-model results from WCRP's CMIP5 *Climate Change-Resilient Agriculture and Agroforestry: Ecosystem Services and Sustainability, Climate Change Management* ed P Castro, A M Azul, W Leal Filho and U M Azeiteiro (Cham: Springer International Publishing) pp 137–56

GDAL/OGR contributors 2019 *GDAL/OGR Geospatial Data Abstraction Software Library. A Translator Library for Raster and Vector Geospatial Data Formats* (Open Source Geospatial Foundation)

Hazeu G, Elbersen B, Andersen E, Baruth B, Diepen K and Metzger M 2010 A biophysical typology in agri-environmental modelling *Environmental and Agricultural Modeling* ed F M Brouwer and M K Ittersum (Dordrecht: Springer) pp 159–87

Hengl T *et al* 2017 SoilGrids250m: global gridded soil information based on machine learning *PLoS One* **12** e0169748

Janík T and Romportl D 2016 Comparative landscape typology of the Bohemian and Bavarian Forest National Parks *Eur. J. Environ. Sci.* **6** 114–8

Jungandreas A, Roilo S, Strauch M, Václavík T, Volk M and Cord A F 2022 Response of endangered bird species to land-use changes in an agricultural landscape in Germany *Reg. Environ. Change* **22** 19

Kassambara A 2017 *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning* (STHDA)

Kohonen T 2001 *Self-Organizing Maps* 3rd edn (*Springer Series in Information Sciences*) (Berlin: Springer) p 502

Lesiv M *et al* 2019 Estimating the global distribution of field size using crowdsourcing *Glob. Change Biol.* **25** 174–86

Levers C, Schneider M, Prishchepov A V, Estel S and Kuemmerle T 2018 Spatial variation in determinants of agricultural land abandonment in Europe *Sci. Total Environ.* **644** 95–111

Li Y, Wright A, Liu H, Wang J, Wang G, Wu Y and Dai L 2019 Land use pattern, irrigation, and fertilization effects of rice-wheat rotation on water quality of ponds by using self-organizing map in agricultural watersheds *Agric. Ecosyst. Environ.* **272** 155–64

Lomba A, Alves P, Jongman R H G and McCracken D I 2015 Reconciling nature conservation and traditional farming practices: a spatially explicit framework to assess the extent of High Nature Value farmlands in the European countryside *Ecol. Evol.* **5** 1031–44

Malek Ž and Verburg P 2017 Mediterranean land systems: representing diversity and intensity of complex land systems in a dynamic region *Landsc. Urban Plan.* **165** 102–16

Mariette J and Villa-Vialaneix N 2016 Aggregating self-organizing maps with topology preservation *Advances in Self-Organizing Maps and Learning Vector Quantization, Advances in Intelligent Systems and Computing* ed E Merényi, M J Mendenhall and P O'Driscoll (Cham: Springer International Publishing) pp 27–37

Metzger M J, Bunce R G H, Jongman R H G, Sayre R, Trabucco A and Zomer R 2013 A high-resolution bioclimate map of the world: a unifying framework for global biodiversity research and monitoring *Glob. Ecol. Biogeogr.* **22** 630–8

Meyer M A and Früh-Müller A 2020 Patterns and drivers of recent agricultural land-use change in Southern Germany *Land Use Policy* **99** 104959

Meyfroidt P *et al* 2018 Middle-range theories of land system change *Glob. Environ. Change* **53** 52–67

Mücher C A, Klijn J A, Wascher D M and Schaminée J H J 2010 A new European Landscape Classification (LANMAP): a transparent, flexible and user-oriented methodology to distinguish landscapes *Ecol. Indic.* **10** 87–103

Oberlack C *et al* 2019 Archetype analysis in sustainability research: meanings, motivations, and evidence-based policy making *Ecol. Soc.* **24** 26

Park Y-S, Kwon Y-S, Hwang S-J and Park S 2014 Characterizing effects of landscape and morphometric factors on water quality of reservoirs using a self-organizing map *Environ. Model. Softw.* **55** 214–21

PBL, 2012. Greening the CAP: an analysis of the effects of the European Commission's proposals for the Common Agricultural Policy 2014–2020 *PBL Neth. Environ. Assess. Agency* (available at: www.pbl.nl/en/publications/greening-the-cap-an-analysis-of-the-effects-of-the-european-commission%E2%80%99s-proposals-for-the-common-agricultural) (Accessed 3 July 2022)

Petrakieva L and Fyfe C 2003 Bagging and bumping self organising maps *Comput. Sci. Inf. Syst.* **9** 69

Plieninger T, Draux H, Fagerholm N, Bieling C, Bürgi M, Kizos T, Kuemmerle T, Primdahl J and Verburg P H 2016 The driving forces of landscape change in Europe: a systematic review of the evidence *Land Use Policy* **57** 204–14

R Core Team 2019 *R: A Language and Environment for Statistical Computing* (Vienna: R Foundation for Statistical Computing)

Rega C, Short C, Pérez-Soba M and Luisa Paracchini M 2020 A classification of European agricultural land using an energy-based intensity indicator and detailed crop description *Landsc. Urban Plan.* **198** 103793

Rocha J, Malmborg K, Gordon L, Brauman K and DeClerck F 2020 Mapping social-ecological systems archetypes *Environ. Res. Lett.* **15** 034017

Seppelt R, Arndt C, Beckmann M, Martin E A and Hertel T W 2020 Deciphering the biodiversity–production mutualism in the global food security debate *Trends Ecol. Evol.* **35** 1011–20

Sietz D, Frey U, Roggero M, Gong Y, Magliocca N, Tan R, Janssen P and Václavík T 2019 Archetype analysis in sustainability research: methodological portfolio and analytical frontiers *Ecol. Soc.* **24** art34

Sroka W, Dudek M, Wojewodzic T and Król K 2019 Generational changes in agriculture: the influence of farm characteristics and socio-economic factors *Agriculture* **9** 264

Stoate C, Báldi A, Beja P, Boatman N D, Herzon I, van Doorn A, de Snoo G R, Rakosy L and Ramwell C 2009 Ecological impacts of early 21st century agricultural change in Europe—a review *J. Environ. Manage.* **91** 22–46

Templ B *et al* 2018 Pan European Phenological database (PEP725): a single point of access for European data *Int. J. Biometeorol.* **62** 1109–13

Trnka M *et al* 2016 Changing regional weather crop yield relationships across Europe between 1901 and 2012 *Clim. Res.* **70** 195–214

Václavík T, Langerwisch F, Cotter M, Fick J, Häuser I, Hotes S, Kamp J, Settele J, Spangenberg J H and Seppelt R 2016 Investigating potential transferability of place-based research in land system science *Environ. Res. Lett.* **11** 095002

Václavík T, Lautenbach S, Kuemmerle T and Seppelt R 2013 Mapping global land system archetypes *Glob. Environ. Change* **23** 1637–47

van Vliet J, de Groot H L F, Rietveld P and Verburg P H 2015 Manifestations and underlying drivers of agricultural land use change in Europe *Landsc. Urban Plan.* **133** 24–36

Vesanto J and Alhoniemi E 2000 Clustering of the self-organizing map *IEEE Trans. Neural Netw.* **11** 586–600

Wehrens R and Kruisselbrink J 2018 Flexible self-organizing maps in kohonen 3.0 *J. Stat. Softw.* **87** 1–18

Will M, Dressler G, Kreuer D, Thulke H-H, Grêt-Regamey A and Müller B 2021 How to make socio-environmental modelling more useful to support policy and management? *People Nat.* **3** 560–72

Wohner C, Ohnemus T, Zacharias S, Mollenhauer H, Ellis E C, Klug H, Shibata H and Mirtl M 2021 Assessing the biogeographical and socio-ecological representativeness of the ILTER site network *Ecol. Indic.* **127** 107785

Ziv G *et al* 2020 BESTMAP: behavioural, ecological and socio-economic tools for modelling agricultural policy *Res. Ideas Outcomes* **6** e52052