# Drivers of global irrigation expansion: the role of discrete global grid choice

Sophie Wagner [1], Fabian Stenzel [2,3], Tobias Krueger [4], and Jana de Wiljes [5,6]

[1]University of Potsdam, August-Bebel-Straße 89, 14482 Potsdam-Griebnitzsee
[2]Potsdam Institute for Climate Impact Research (PIK), Member of the Leibniz Association, P.O. Box 60 12 03, 14412 Potsdam, Germany, stenzel@pik-potsdam.de
[3]Stockholm Resilience Centre, Stockholm University, Sweden
[4]Geography Department & IRI THESys, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, tobias.krueger@hu-berlin.de
[5]Institute for Mathematics, Ilmenau University of Technology, Weimarer, 98693 Ilmenau, jana.de-wiljes@tu-ilmenau.de
[6]School of Engineering Sciences, Department of Computational Engineering, Yliopistonkatu 34 53850 Lappeenranta, Finland LUT University, Finland

**Correspondence:** Sophie Wagner (sopwagner@uni-potsdam.de)

**Abstract.** Global statistical irrigation modeling relies on geospatial data and traditionally adopts a discrete global grid based on longitude-latitude reference. However, this system introduces area distortion, which may lead to biased results. We propose using the ISEA3H geodesic grid based on hexagonal cells, enabling efficient and distortion-free representation of spherical data. To understand the impact of discrete global grid choice, we employ a non-parametric statistical framework, utilizing random forest methods, to identify the main drivers of historical global irrigation expansion using, among other data, outputs from the global dynamic vegetation model LPJmL.

Irrigation is critical for food security amidst growing populations, changing consumption patterns, and climate change. It significantly boosts crop yields but also alters the water cycle and global water resources. Understanding past irrigation expansion and its drivers is vital for global change research, resource assessment, and predicting future trends.

We compare predictive accuracy, simulated irrigation patterns and identification of irrigation drivers between the two grid systems. Using the ISEA3H geodesic grid system increases the predictive accuracy by up to 28% compared to the longitude-latitude grid. The model identifies population density, potential productivity increase, evaporation, precipitation, and water discharge as key drivers of historical global irrigation expansion. GDP per capita also shows some influence.

We conclude that the geodesic discrete global grid system significantly affects predicted irrigation patterns and identification of drivers, and thus has the potential to enhance statistical modeling, which warrants further exploration in future research across related fields. This analysis lays the foundation for comprehending historical global irrigation expansion.

## 1 Introduction

About 80% of data being produced is of geospatial nature (Hahmann and Burghardt, 2013). While the construction of maps and the referencing of locations on the Earth's surface has a very long history, it is becoming increasingly important to find efficient

ways to process, integrate and analyze geospatial data to solve problems in times of globalization. To that end, geographic grid systems are used to project the geographic space into a mathematical space where algorithms and statistical methods can be applied.

The most widely used grid system is the geographic coordinate system (using latitude and longitude lines), which dates back to the third-century BCE (McPhail, 2011; Ware et al., 2020). A great advantage of this system is that it can be stored compactly and used easily for computations (Ware et al., 2020). Unfortunately, when portrayed on a sphere, grids based on geographic coordinates suffer from cell area distortion due to the converging lines of equal longitude. In the context of global statistical modelling, this ultimately results in oversampling of the northernmost regions.

A great number of alternative Discrete Global Grid Systems (DGGS) have emerged to fulfill the needs of different research fields and modelling strategies (Goodchild, 1994). Since these grid systems offer significant benefits for Big Data and Digital Earth research, there have been numerous advancements, new implementations, and example applications in various fields (Sahr, 2011; Jendryke and McClure, 2019; Sirdeshmukh et al., 2019; Bousquin, 2021, e.g.). Today, standard Earth grid systems, including DGGS, are documented by the Open Geospatial Consortium (OGC) and the International Organization for Standardization (ISO) (ISO, 2021; Purss, 2015). Even though current implementations might not yet fully comply with all these standards (Bondaruk et al., 2020), there is a clear benefit to starting the integration of these data structures.

In this paper, we propose to use a global DGGS based on a hexagonal tessellation of the Earth's surface in the context of modelling global historical irrigation expansion. This grid was introduced by Sahr et al. (2003) and has gained large popularity in many research contexts. Two recent examples of open-source DGGS libraries that are based on hexagonal grid structures are the H3 system, developed by Uber (2022) and DGGRID (Barnes and Sahr, 2017). Mechenich and Zliobaite (2023) recently presented the Eco-ISEA3H database that consists of global spatial data characterizing climate, geology, land cover, physical and human geography, and the geographic ranges of nearly 900 large mammalian species. In contrast to grid cells induced by the longitude-latitude graticule, hexagonal cells are able to cover almost the entire surface of the Earth without suffering from area distortion. That way, all regions have the same influence in a statistical model.

Recently, hexagonal mesh grids have gained popularity among hydrologists (Li et al., 2022). A group of hydrological functions on hexagonal meshes, such as flow direction and accumulation, stream networks, or watershed boundary extraction, were explored by Liao et al. (2020). The authors show that their algorithm's performance is better when considering the hexagonal-mesh-based output compared to the traditional square-mesh-based output. Wang et al. (2020) studied valley networks and model valley lines based on hexagonal grids. Compared to traditional square grids, the study shows that using the hexagonal grid leads to a higher location accuracy. In another study, Wright (2019) developed a regular hierarchical surface model where hydrological computation was generalized on hexagonal and triangular grids. Additionally, there has been an increasing interest in managing geospatial data and developing models to solve real-world problems by using the open-source DGGRID library (Hojati and Robertson, 2020; Li et al., 2021; Chaudhuri et al., 2021; Robertson et al., 2020; Li et al., 2022).

Our study aims to predict global historical irrigation patterns. Since both rainfed and irrigated grid cells are equally important for our analysis, addressing the issue of area distortion is crucial. To the best of our knowledge, we are the first to utilize the Icosahedral Equal Area aperture 3 Hexagon geodesic Discrete Global Grid System (ISEA3H DGGS) in this context. This

grid system enables us to directly analyse global irrigation patterns without area distortion. Additionally, the grid system has potential for improving the mapping of spatial clusters and neighborhood patterns, as each grid cell has a unique set of neighbors.

The second aim of this paper is to contribute to the literature on global irrigation expansion. Irrigation is crucial to ensure the world's food security. A growing human population, shifting consumption patterns and climate change increase the pressure on agricultural production (Foley et al., 2011). To meet the growing human food demand, irrigation has rapidly increased over the last century as it increases crop yields (Siebert et al., 2015). In the year 2000, approximately $40\%$ of the global food production was harvested on irrigated land, utilizing only 20% of the total farming area (Schultz et al., 2005). To achieve this agricultural intensification, a large amount of fresh water is needed. Consequently, irrigation alters the hydrological cycle significantly (Zohaib and Choi, 2020). At a global scale, irrigation is responsible for about 60% of total fresh water withdrawals and 80% of total fresh water consumption (Döll et al., 2014; Siebert et al., 2015). It is therefore important to understand the past evolution of irrigation expansion and its main drivers for global change research, the assessment of resources and for predicting future developments.

There have been a few studies on the drivers of global irrigation in previous years. Neumann et al. (2011) investigated the global irrigation pattern in the year 2000. Using a multilevel approach, they modeled irrigation as a function of biophysical and socioeconomic factors. Their results show that biophysical factors have significant influence on irrigation. Additionally, the authors provide suggestive evidence that socioeconomic factors play a role for irrigation. However, it is emphasised that the model suffers from uncertainty due to the lack of spatially explicit socioeconomic information and the possibility of external influences, such as public investments. While our model also faces these limitations, we are able to extend the analysis by including a historical dimension.

Puy et al. (2020) investigated uncertainties in published projections of global irrigation expansion for the year 2050. By comparing different projected estimates of irrigated area to a simple model predicting irrigated area as a function of only population size, constrained by water and land availability, taking into account parametric and model uncertainties, the authors postulate, that current models underestimate future irrigated areas. Other recent studies developed global irrigation maps, mostly using a combination of remote sensing, machine learning and climate data (Meier et al., 2018; Salmon et al., 2015; Nagaraj et al., 2021).

We contribute to the literature on global irrigation expansion by investigating the drivers of the historical expansion between 1902 and 2000, using a novel non-parametric statistical model. We distinguish between the factors that influence the probability of a grid cell being irrigated, i.e. the decision to irrigate instead of remaining rainfed, and the irrigation intensity, once a grid cell is irrigated. We employ a stacked random forest framework to assess the quality between statistical irrigation models based on two different grid systems. Our main focus is on the influence of the grid the data is presented in.

## 2 Data

Our first objective in this study is to analyze the choice of discrete global grid system for modelling the historical evolution of global irrigation expansion. To achieve this, we focus on the entire global land surface, excluding Antarctica. We consider data from 1902 to 2005 to comprehensively capture the historical evolution of irrigation expansion over the past century.

Our analysis builds on a data set that consists of a simulation output from the *Lund-Potsdam-Jena managed Land* (LPJml) model (Sitch et al., 2003; Bondeau et al., 2007) and historical economic data from the *Maddison Project Database* (Inklaar et al., 2018).

LPJmL is a process-based dynamical global vegetation, hydrology, and crop model simulating natural and managed vegetation growth based on soil, climate, and management input at a daily resolution and at a global $0.5° \times 0.5°$ spatial grid scale, resulting in a total amount of 67420 terrestrial grid cells per time unit in each variable (Schaphoff et al., 2018).

We prescribe an agricultural land use data set based on the *History Database of the Global Environment* (HYDE) (Klein Goldewijk et al., 2017) with additional assumptions on irrigation systems and extent of areas equipped for irrigation by Jägermeyr et al. (2015) based on the *Global Historical Irrigation Data Set* (HID). One advantage of the HID is, that the evolution of land irrigation was implemented using official land use data and is therefore independent of socioeconomic information, such as gross domestic product or population density (Siebert et al., 2015). Hence, the relationship of irrigation and socioeconomic variables can safely be analyzed. As climate input, the *Climatic Research Unit Timeseries* (Harris et al., 2014) is used. Whether a crop actually needs irrigation is internally decided by the LPJmL simulation based on biophysical constraints, and constrained by surface water availability (Schaphoff et al., 2018).

From the simulation, we obtain the direct output variables precipitation, evaporation, discharge, crop yield, and the actually irrigated fraction for each grid cell. Additionally the median potential increase in crop yield productivity is derived. This is estimated from two separate synthetic simulations, where potential yields for each crop and grid cell are compared with and without irrigation. The variables are aggregated annually for each grid cell to obtain a time series for the years 1901 to 2005.

The LPJmL data are complemented by the Maddison Project database on the historical performance of the world economy (Inklaar et al., 2018; Bolt and Zanden, 2014). Of particular interest is the gross domestic product (GDP) per capita time series, consisting of estimates of comparative levels of real GDP per capita in recent time periods, combined with long-term time series growth of GDP per capita. Even though the Maddison Project database yields state of the art historical economic data, there are many countries without an estimation of GDP per capita in the time period 1900 to 1960, leading to missing data.

### 2.1 Variables

We use the fraction of a grid cell that is actually irrigated as the dependent variable. ~~If~~ This fraction is a continuous variable between 0 and 1, with 1 meaning, a grid cell is fully irrigated, ~~the variable is assigned a value of "1"; if~~ and 0 meaning that none of the area is irrigated~~, the variable is assigned a value of "0."~~ . It is important to note that since grid cells in the standard longitude-latitude grid change in area relative to latitude, irrigation fraction values between grid cells are not directly comparable. Figure A4 displays the global irrigation fraction map, based on HID data from 2000.

The selection of potential drivers of irrigation expansion was led by existing literature and data availability (see Table 1). We consider the following variables for explaining irrigation fraction: population density, precipitation, discharge, evaporation, potential yield increase through irrigation, and GDP per capita.

The GDP per capita data is available at national level and broken down to a $0.5° \times 0.5°$ grid scale, by assigning the country's value to all grid cells in a country. Since there are observations missing, especially in the earlier time periods, the variable is split up into the categories "high income", "upper middle income", "lower middle income", "low income" and "missing", following the methodology of Hastie et al. (2009) and the World Bank's classification of GDP per capita from 2011 (World Bank, 2011). The classification can be found in the supplementary material (Table A2). That way, we treat the missing values as an additional category and are able to include all observations in our analysis.

We report the pairwise Pearson correlation coefficients (supplementary material, Table A3) and variance inflation factors (supplementary material, Table A4) to investigate multicollinearity between the continuous predictor variables, following the methodology in Rufin et al. (2018). We find that all pairwise Pearson correlation coefficients are below the threshold of 0.7 (Dormann et al., 2013), except for the pair "Precipitation" and "Evaporation", where we find a value of 0.72. The variance inflation factors are below the tight threshold of 5, indicating that the predictor variables are sufficiently independent for our analysis (James et al., 2013).

Table 1: Potential predictors and hypotheses

| Predictor Variable | Hypothesis | Supporting Literature |
|---|---|---|
| Precipitation (mm/year) | Irrigation requirements increase in cropland regions where precipitation levels are declining. | (Neumann et al. (2011)), (Döll and Siebert (2002)), (Siebert et al. (2015)) |
| Discharge ($hm^3$/year) | Surface water availability allows for irrigation water withdrawals. | (Neumann et al. (2011)), (Gerten et al. (2008)) |
| Evaporation (mm/year) | High evaporation leads to an increasing demand of water and therefore increases the probability of irrigation. | (Neumann et al. (2011)), (Rufin et al. (2018)) |
| Median Increase in Productivity (% of $\Delta$ gC/$m^2$) | If the potential increase in agricultural productivity is large, the corresponding area is more likely to receive irrigation. | (FAO and of the United Nations (2011)), (Sauer et al. (2010)) |

| Predictor Variable | Hypothesis | Supporting Literature |
|---|---|---|
| Population Density (cap/$m^2$) | Intensive irrigation occurs under high population densities. The rapidly growing world population increases the demand for food and, therefore, leads to an expansion or intensification of agriculture globally but also around high-density centres. | (Neumann et al. (2011)), (Rufin et al. (2018)), (Boretti and Rosa (2019)), (Sauer et al. (2010)) |
| GDP ($ US /cap) | A high GDP per capita leads to a higher probability of irrigation, since farmers can afford irrigation systems or are more likely to receive subsidies. GDP is also highly correlated with government effectiveness and hence serves as a proxy. A high national government effectiveness strengthens irrigation infrastructure. | (Neumann et al. (2011)), (Rufin et al. (2018)), (Boretti and Rosa (2019)), (Sauer et al. (2010)) |

### 2.2 Descriptive statistics

Overall, the irrigated area expanded throughout the study period. The proportion of grid cells with observed irrigation increased from approximately 10% in 1902 to about 31% in 2005. The most significant increases in irrigated area occurred in southeastern Asia, Middle and South America, Central America, and eastern Asia. These statistics align with the findings of Siebert et al. (2015), who investigated areas equipped for irrigation.

Despite the expansion of irrigated land, the data is highly imbalanced: throughout the study period, about 75% of the observed irrigation fractions are zero, whereas only about 25% are non-zero. Figure A2 in the appendix shows the histogram of the irrigation fraction.

The descriptive statistics of irrigation fraction and the potential predictors can be found in Table A1. The dependent variable, irrigation fraction, ranges from zero to 0.922 with a mean of 0.008 in the longitude-latitude grid.

The global temporal evolution of the predictor variables is illustrated in Figure A1 in the appendix. The global mean evaporation has been increasing over the last century as well as the GDP per capita. We also see a slightly increasing trend of the global amount of precipitation. For the remaining variables, there is no clear detectable trend in the global mean. However, it is expected that there are local trends that are not captured in the global mean values.

## 3 Method

### 3.1 Spatial resolution

The latitude-longitude projection yields a world map which appeals to the human eye for its plane appearance but also faces some limitations. The grid cells that are induced by the longitude-latitude graticule are not of the same area. One degree of latitude represents the same horizontal distance anywhere on the Earth's surface. However, because lines of equal longitude are farthest apart at the equator and converge to single points at the geographic poles, the horizontal distance equivalent to one degree of longitude, varies with latitude (Budic et al., 2016). For a simple statistical analysis, this implies that regions nearer

the poles, which are smaller in area yet weighted equally to larger areas nearer the equator, disproportionately contribute to models and thus have a greater influence on the results. This is particularly relevant for land-based analysis like ours, as a significant land mass in the northern hemisphere is located closer to the poles.

The limitations of discrete global grids rooted in the geographic coordinate system have spurred the exploration of various alternatives. One idea, is to weight each grid cell by its area and therefore its relative importance for the statistical model. A more direct approach is to use a discrete grid system that subdivides the Earth's surface into equally sized grid cells, allowing for an efficient identification of patterns, trends and relationships across diverse geographic scales.

Sahr et al. (2003) introduced a class of reference grids based on convex regular polyhedra, called *geodesic Discrete Global Grid Systems* (geodesic DGGS). The underlying idea is to use the topological equivalence of regular polyhedra and the sphere. Based on five design choices, the resulting grid partitions the Earth into equally-sized cells. The first choice involves picking a base polyhedron. The distortion of area tends to be smaller the smaller the faces of the base polyhedron (Sahr et al., 2003). Therefore, in this study we choose the icosahedron as a starting point, as it has the smallest face sizes compared to the other regular polyhedra. The second design choice requires to pick a method of partitioning the surface of the icosahedron. Hexagons have been found in many research fields to be the optimal choice for discrete gridding and location representation (Apte et al., 2013; Uher et al., 2019). One unique property of a hexagonal grid is its uniform adjacency; each cell in a hexagonal grid has six neighbors, all of which share an edge with the cell, and all of which have centers exactly the same distance away from their neighbouring cells. This property is beneficial for all analyses involving neighborhood properties.

Thirdly, one has to decide on the orientation of the base icosahedron relative to the Earth's surface. In other words, the location of the pentagonal cells is required, as they are located at the vertices of the icosahedron. The most common choice is to place the pentagons, such that only one is centered on land (Sahr et al., 2015). This specific orientation is also symmetric about the equator.

In a fourth step, a method for the transformation between the surface of the Earth and the surface of the icosahedron, upon which the hexagonal grid is constructed has to be selected. Our choice is the only known equal-area icosahedral geodesic DGG projection, called the Snyder Icosahedral Equal Area (ISEA) projection (Snyder, 1992).

Lastly, a recursive partitioning method must be picked in order to create different spatial resolutions. Such method is characterized by the ratio of cell areas at a given grid resolution and the next coarser resolution. This ratio is called *aperture*. We will consider aperture 3 hexagonal grid cells, meaning that the increase of the resolution by one, leads to grid cells with an area of a third of the original cell area. Figure A5 in the appendix illustrates the partitioning method.

After making these five basic construction choices, the result can be referred to as an *Icosahedral Snyder Equal Area aperture 3 Hexagon geodesic Discrete Global Grid System* (ISEA3H DGGS).

### 3.2 Data transformation

Our data set is initially organized based on the standard longitude-latitude reference system, with a spatial resolution of $0.5° \times 0.5°$. In this system, the location of a grid cell is determined by the latitude and longitude of its center point. This results in a total of 67,420 land grid cells per year. To assess the impact of different discrete global grid choices and compare between

the standard approach and the distortion-free geodesic alternative, we constrict a geodesic DGG reference framework and transform our observations accordingly.

Utilizing the freely available $R$ package *dggridR* provided by Barnes and Sahr (2017), we generate ISEA3H discrete global grids at resolutions 7, 8 and 9. This translates to hexagonal grid structures, where the centers of grid cells are spaced 160 km, 95 km, and 55 km apart. Following the transformation, we obtain 7,383, 65,612, and 196,832 terrestrial grid cells per year, respectively. We use three different resolutions to explore the model's sensitivity to grid resolution and ensure the robustness of our findings. We start with resolution 7 due to its efficiency. The resolution 8 grid contains a total of 65,612 grid cells, comparable to the longitude-latitude grid's 67,420 cells. Meanwhile, the resolution 9 grid features average grid cell sizes similar to those in the longitude-latitude grid, with cells spanning 2,591.402 $km^2$, compared to the longitude-latitude grid's average of 2,171.119 $km^2$.

The original data are projected into the hexagonal cells. Depending on the degree of latitude, different amounts of cell centers of the original grid end up in each hexagon. The center counts at resolution 7 are visualized in Figure 1.

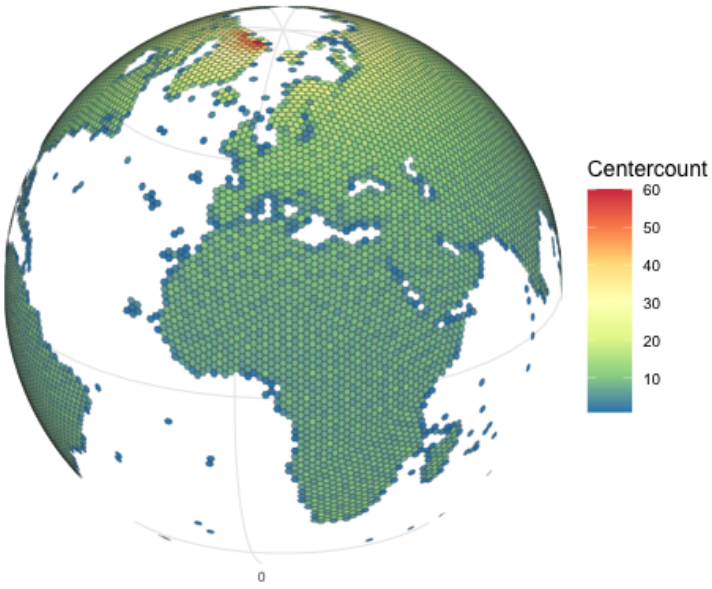



**Figure 1.** Number of grid cell centers of the longitude-latitude grid that fall into each hexagonal grid cell of the ISEA3H grid at resolution 7. The colour pattern shows that ~~in~~ closer to the ~~northern (and southern) parts~~poles, more grid cell centers fall into each hexagonal cell compared to the areas around the equator.

After mapping the original cell centers into the ISEA3H grids, the mean of all observations within each hexagonal cell is taken as the new value in the transformed data set.

### 3.3 Area weights

Instead of directly converting the data into the area-preserving ISEA3H grid system, one could alternatively adjust the observations from the longitude-latitude grid by weighting them according to their cell size to account for the area distortion. The weight $w_i$ of grid cell $i$ is then defined as $w_i = \frac{\text{area of cell } i}{\text{maximum cell area}}$. By including these area weights, observations from larger cells have a greater influence on the statistical model than cells from smaller cells. We use this approach as a comparison to better assess the effectiveness of using traditional area weights to address area distortion, in contrast to the equal-area ISEA3H grid system.

### 3.4 Random forest

We model the observed variation of irrigation fraction with a set of biophysical and socioeconomic predictor variables using a random forest framework.

A random forest consists of a set of individual decision trees that operate as an ensemble. The method was introduced by Breiman (2001) and is now a widely used machine learning technique, because it tends to have high prediction power with little tuning of its parameters. A random forest captures non-linear relationships between the predictor variables and the outcome, is able to deal with imbalanced data, and estimates of variable importance are readily available (Strobl et al., 2009).

Depending on the response variable, the decision trees of the random forest perform either classification or regression, based on a recursive partitioning method. At each step, a decision tree finds the optimal split that minimises "impurity", until a stopping criterion is met. Impurity serves as a metric for the homogeneity of the class labels at a particular node within the decision tree. Various methods exist to define the impurity measure. Following Wright and Ziegler (2017), we use the estimated response variance for regression trees and the Gini-index for classification trees as measures for impurity. Please find the precise steps in Algorithm 1. Ultimately, the recursive partitioning method repeatedly splits the data into potentially high-dimensional rectangular partitions of the predictor space, choosing those for which the response data are relatively homogeneous (Strobl et al., 2009).

A random forest typically consists of several hundred or thousands of trees and combines the results of their predictions (Strobl et al., 2009). These trees are constructed using bootstrapped samples from the training data, with each sample containing, on average, 63.2% unique observations (Breiman, 2001), known as in-bag samples. Samples not selected are called out-of-bag (OOB) samples and are used to estimate the prediction accuracy, also called *OOB error*. These error estimates provide an accurate measurement of the generalization error as they are similar to the results obtained through $K$-fold cross-validation (Wolpert and Macready, 1996). However, the OOB error can be sensitive to the number of random predictors used at each split ($mtry$) and the number of trees ($ntree$) in the random forest (Huang and Boutros, 2016). Generally, the accuracy increases as the number of trees increases. However, the accuracy may level off at a certain number of trees, depending on the specific learning task (Oshiro et al., 2012). The parameter $mtry$ has been found to have a high influence on prediction accuracy

and should be selected carefully (Huang and Boutros, 2016; Bernard et al., 2009; Probst et al., 2019). We focus on $ntree$ and $mtry$ as tuning parameters to achieve a high performing random forest model.

The step-by-step process of building a classification and regression random forest follows Algorithm 1. To cope with the

**Algorithm 1** Random Forest

Given a data set $\{(x_i, y_i) : i = 1, ..., n\}$, where $y_i$ is the $i$th observed dependent variable and $x_i = (X_1, ..., X_p)$ is a $p$-dimensional predictor vector.

**Step 1.** Draw a number of $ntree$ bootstrap samples sets from the training data set. Each sample is the same size as the training data set. The number $ntree$ is a tuning parameter, also referred to as the number of trees in the forest.

**Step 2.** At each node split a random number of $mtry$ predictors out of all $P$ predictors are considered, i.e. $X_i, i = 1, ..., mtry$ with $mtry < P$. The number $mtry$ is another tuning parameter.

**Step 3.** Predictor $j$ splits the observations $\{y_i\}, i = 1, ..., n$ into the most uniform binary regions $R_l := \{X | X_j \leq c\}$ and $R_r := \{X | X_j > c\}$ according to the following impurity measures:

- (Regression) weighted residual sums of squares

$$min_{j,c}\Big(p(R_l) \sum_{j:y_j \in R_l} (y_j - \bar{y}_{R_l})^2 + p(R_r) \sum_{j:y_j \in R_r} (y_j - \bar{y}_{R_r})^2\Big), \tag{1}$$

  where $\bar{y}_{R_l}$ and $n_l$ are the mean and number of observations in region $R_l$, $\bar{y}_{R_r}$ and $n_r$ are the mean and number of observations in region $R_r$ and $p(R_k) = n_k/n$ is the proportion of observations in Region $k \in \{l, r\}$.

- (Classification) Gini impurity

$$min_{j,c}\Big(n_l \hat{p}_l (1 - \hat{p}_l) + n_r \hat{p}_r (1 - \hat{p_r})\Big), \tag{2}$$

  where $\hat{p}_k$ is the proportion of sample points that were sent to node $k \in \{l, r\}$ from the previous node.

**Step 4.** Repeat steps 2-3 until each terminal node reaches the predefined minimum number of observations $min.node.size$.

**Output.** The algorithm forms a partition of the data into $M$ regions $R_1, ..., R_M$, and model the response as a constant $r_m$, i.e.:

$$f_{RF}(x) = \sum_{m=1}^{M} r_m I(x \in R_m). \tag{3}$$

imbalance of our dependent variable, we train two random forests and construct a hurdle model. A classification random forest predicts whether a grid cell is irrigated or not, while a regression random forest predicts the magnitude of irrigation. These models are then combined to create a stacked final model that predicts the irrigation fraction based on the available predictors. This approach effectively handles the zero-inflated distribution of the irrigation fraction.

We use the freely available $R$ package *ranger*, developed by Wright and Ziegler (2017) for the training and validation of the random forests.

### 3.4.1 Parameter tuning and model setup

We use cross-validation (CV) to tune our random forest models and determine the values for *ntree* (number of decision trees) and *mtry* (number of predictors to be considered at each split) that maximize predictive accuracy. Data from 1902 to 1999 serve as the training sample, data from 2001 to 2005 are used for validation, and data from 2000 act as test sample. Due to computational constraints, we apply a sub-sampling routine to identify our model parameter values efficiently. For the classification random forests, we draw a balanced sample of 10% in each CV fold, consisting of 50% irrigated and 50% rainfed grid cells. This is achieved using random over- and under-sampling methods from the R package ROSE, provided by Lunardon et al. (2014). For the regression random forest, all irrigated grid cells are used for training.

We use the OOB error and the validation error as accuracy measures. We set the minimal number of data points at each terminal node *min.node.size* to 10, serving as a stopping criterion. For the parameter *ntree*, we consider the values 50, 300, 500, 800, 1,000, 2,000, 3,000, 4,000 and 5,000. For *mtry* we test all values between 1 and 5 at 0.5 increments for both the classification and the regression random forest. We conduct 50-fold CV to train the classification and regression random forests separately for each grid choice.

The resulting accuracy for each forest and each tuning parameter value can be seen in Figures A6 and A7. Taking the OOB error and the validation error into account, we choose $ntree =$ 1,000 and $mtry =$ 1.5 for the classification random forest and $ntree =$ 4,000 and $mtry =$ 5 for the regression random forest for the longitude-latitude grid. In the ISEA3H grid we set $ntree =$ 1000 and $mtry =$ 5 for the classification random forest and $ntree =$ 4000 and $mtry =$ 5 for the regression random forest.

After setting the tuning parameters, we evaluate the prediction accuracy of the stacked random forest model on the test data. The final model prediction is obtained by multiplying the predictions from the classification random forest with those from the regression random forest.

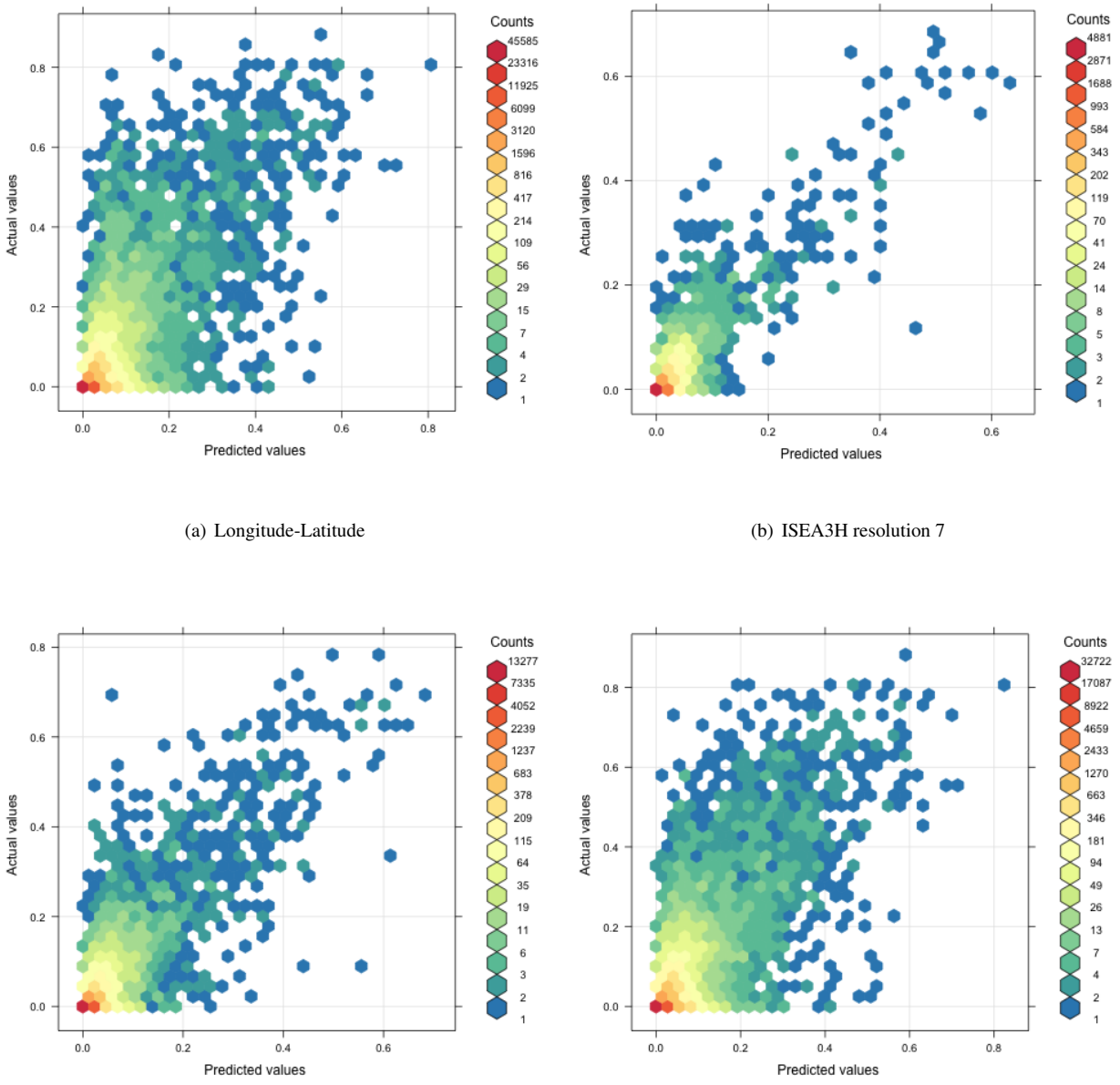
We then compare the final prediction results for models build on the original longitude-latitude grid, the longitude-latitude grid using area weighting, and the ISEA3H grids at resolutions 7, 8 and 9.

## 4 Results and discussion

### 4.1 Grid choice

We compare the longitude-latitude grid to the ISEA3H grids based on their predictive power and their ability to identify the drivers of the global irrigation expansion. To get a first intuition about differences in predictive power, we create binned scatterplots of the predicted irrigation fraction of the test data against the observed values of irrigation fraction for all grid choices. In that way, the 45 degree line mechanically indicates correctly predicted irrigation fraction values. Figure 2 shows the results. The comparison suggests that the ISEA3H grid models at resolution 7 and 8 have a higher prediction accuracy, since the point values scatter more closely around the 45 degree line. ~~However, there~~ This could be due to the fact that at

resolutions 7 and 8, the grid cells are larger and therefore the value of the dependent variable less nuanced. There is no clear visual difference between the predictive accuracy of the longitude-latitude grid and the ISEA3H grid at resolution 9.




(a) Longitude-Latitude




(b) ISEA3H resolution 7




(c) ISEA3H resolution 8




(d) ISEA3H resolution 9

**Figure 2.** Binned scatter plot of predicted vs. observed irrigation fraction values. The prediction is based on the test data.

To further evaluate the difference in predictive accuracy between grid choices, we compute the root mean square error (RMSE) and the normalized root mean square error (NRMSE) as

$$RMSE = \sqrt{1/n \sum_{i=1}^{n} (y_i - \hat{y_i})^2} \tag{4}$$

and

$$NRMSE = \frac{RMSE}{sd(y)}, \tag{5}$$

where $y_i$ is the observed value, $\hat{y}_i$ the prediction and $sd(y)$ the standard deviation over all observed values. The RMSE and NRMSE were calculated for the prediction on the test data and compared between grid choices. We additionally evaluate the NRMSE, after restricting the sample to observations with non-zero irrigation. The outcomes are reported in Table 2. The model with the lower NRMSE is considered the better choice to model irrigation fraction.

To verify the robustness of our result, we calculate the NRMSE by using the mean and the distance between the minimum and the maximum value as standardizing measures.

In order to investigate the robustness of our error measure, we implement a bootstrapping analysis, in which we generate each model in 500 repetitions and predict irrigation fraction using a random 40% sample of the test data in each step. We then calculate the difference in NRMSE values between the longitude-latitude benchmark model and the other specifications. By examining the distribution of these differences, we are able to assess whether observed differences are statistically significant.

Additionally, we include a model based on the longitude-latitude grid with traditional area weights. This allows us to assess the effectiveness of using area weights to address area distortion as compared to the equal-area ISEA3H grid.

Generally, we see that lower errors are observed when using an ISEA3H grid. For all observations, the ISEA3H resolution 7 grid exhibits a 28% reduction in NRMSE compared to the longitude-latitude grid, with a value of 0.484 compared to 0.676. The ISEA3H grids at resolutions 8 and 9 also show improved performance over the longitude-latitude grid, with NRMSE values of 0.577 and 0.645, respectively. The longitude-latitude grid with area weights does not significantly improve the NRMSE compared to the standard longitude-latitude grid.

Focusing on irrigated areas, the ISEA3H resolution 7 grid demonstrates a 29% lower NRMSE (0.503) compared to the longitude-latitude grid (0.702). Similar trends are observed for the ISEA3H grids at resolutions 8 and 9, with NRMSE values of 0.598 and 0.666, respectively. Again, the longitude-latitude grid with area weights shows marginal or no improvement.

This trend remains consistent across all normalization specifications, which emphasizes the comparative performance of the ISEA3H grid choices and their advantages over traditional longitude-latitude grids.

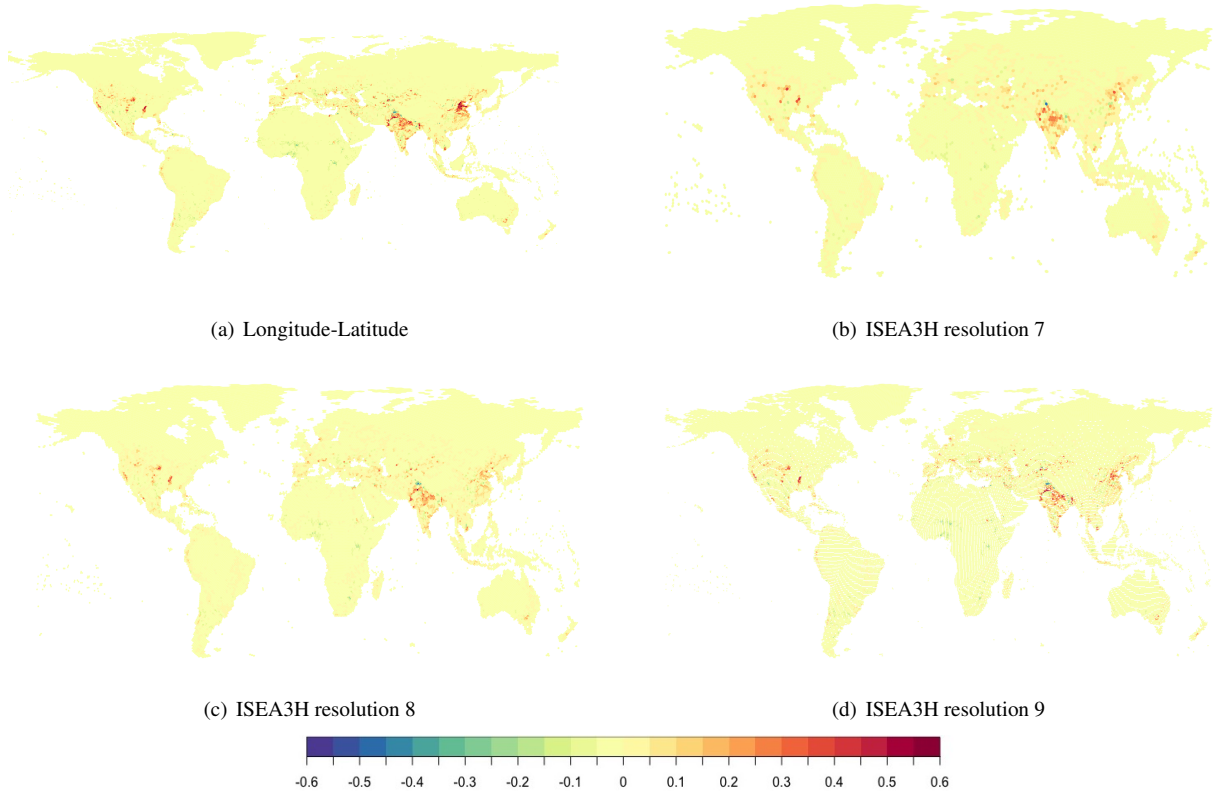
In a next step, we consider predicted irrigation fraction. We evaluate how accurately the models predict high and low values of irrigation fraction across the globe. Figure 3 shows the difference between predicted irrigation fraction and observed irrigation fraction for all grid choices. The computation is based on the test data. The color scale indicates, if the model predicts the irrigation fraction accurately or suffers from under- or over-prediction. Yellow areas are correctly predicted by the model, orange to red areas correspond to under-prediction and green to blue areas indicate over-predicted irrigation fractions.

**Table 2.** Normalized root mean square error comparison between grid choices

| | Longitude-Latitude | ISEA3H resolution 7 | Reduction in NRMSE (%) | Longitude-Latitude area weights | ISEA3H grid resolution 8 | ISEA3H resolution 9 |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| **A. All observations** | | | | | | |
| Mean | 0.0156 | 0.0168 | | 0.0156 | 0.0177 | 0.0183 |
| Mean (prediction) | 0.0162 | | | 0.0161 | 0.0171 | 0.0191 |
| SD | 0.0604 | 0.0525 | | 0.0604 | 0.0591 | 0.0653 |
| SD (prediction) | 0.0396 | | | 0.0396 | 0.0432 | 0.0454 |
| RMSE normalized with: | | | | | | |
| SD | 0.676 | 0.484*** | 28 | 0.676 | 0.577*** | 0.645*** |
| Mean | 2.618 | 1.508*** | 42 | 2.620 | 1.928*** | 2.297*** |
| Max-Min | 0.047 | 0.037*** | 21 | 0.047 | 0.044** | 0.046* |
| **B. Non-zero observations** | | | | | | |
| Mean | 0.0507 | 0.0337 | | 0.0507 | 0.0413 | 0.0511 |
| Mean (predition) | 0.0409 | | | 0.0408 | 0.0154 | 0.0434 |
| SD | 0.1005 | 0.0703 | | 0.1005 | 0.0847 | 0.1010 |
| SD (prediction) | 0.0619 | | | 0.0618 | 0.0401 | 0.0667 |
| RMSE normalized with: | | | | | | |
| SD | 0.702 | 0.503*** | 29 | 0.703 | 0.598*** | 0.666*** |
| Mean | 1.390 | 1.050*** | 24 | 1.391 | 1.228*** | 1.317*** |
| Max-Min | 0.081 | 0.052*** | 36 | 0.081 | 0.065*** | 0.077*** |
| $R^2$ | 0.0694 | 0.7719 | | 0.5527 | 0.6946 | 0.5889 |

Notes: Column (1) shows the mean and standard deviation of the irrigation fraction, and the NRMSE values of the longitude-latitude grid choice. Column (2) provides the same for the ISEA3H grid resolution 7 choice. In column (3) the reduction in NRMSE is documented in percent and in comparison to the longitude-latitude grid. Column (4) presents the result for a model based on the longitude-latitude grid with additional area weights and columns (5) and (6) provide the results for the ISEA3H resolution 8 and 9 grids. Panel A. includes all observations and gives the overall NRMSE estimates. In Panel B. only irrigated areas are included. The NRMSE values provide insight to how the models perform on actually irrigated terrain. $*$,$**$ and $***$ indicate 10%, 5% and 1% significance for the t-test of difference in bootstrapped mean NRMSE values with 500 repetitions, comparing the ISEA3H models (columns 2, 5, and 6) with the longitude-latitude model (column 1).

Considering the longitude-latitude grid, we see that, irrigation is under-predicted in some areas in India and East Asia and also in few areas in North and South America and Europe. Except for very few parts in India, Central-Africa and North America, we do not see any over-prediction of irrigation. Looking at the ISEA3H grids, we find that the same areas in India and East Asia are slightly over-predicted as well as some areas in the United States and Europe. Only few areas are under-predicted in India, East-Asia, and Central-Africa. Comparing both grid systems, we find that the ISEA3H grids are closer to the original irrigation pattern in all areas. Especially, the highly irrigated areas in east Asia are better captured by the ISEA3H grid models and we also see less over-prediction in European areas. The maps indicate that the ISEA3H grid system is the better choice in predicting the global irrigation fraction pattern.




(a) Longitude-Latitude

(b) ISEA3H resolution 7

(c) ISEA3H resolution 8

(d) ISEA3H resolution 9

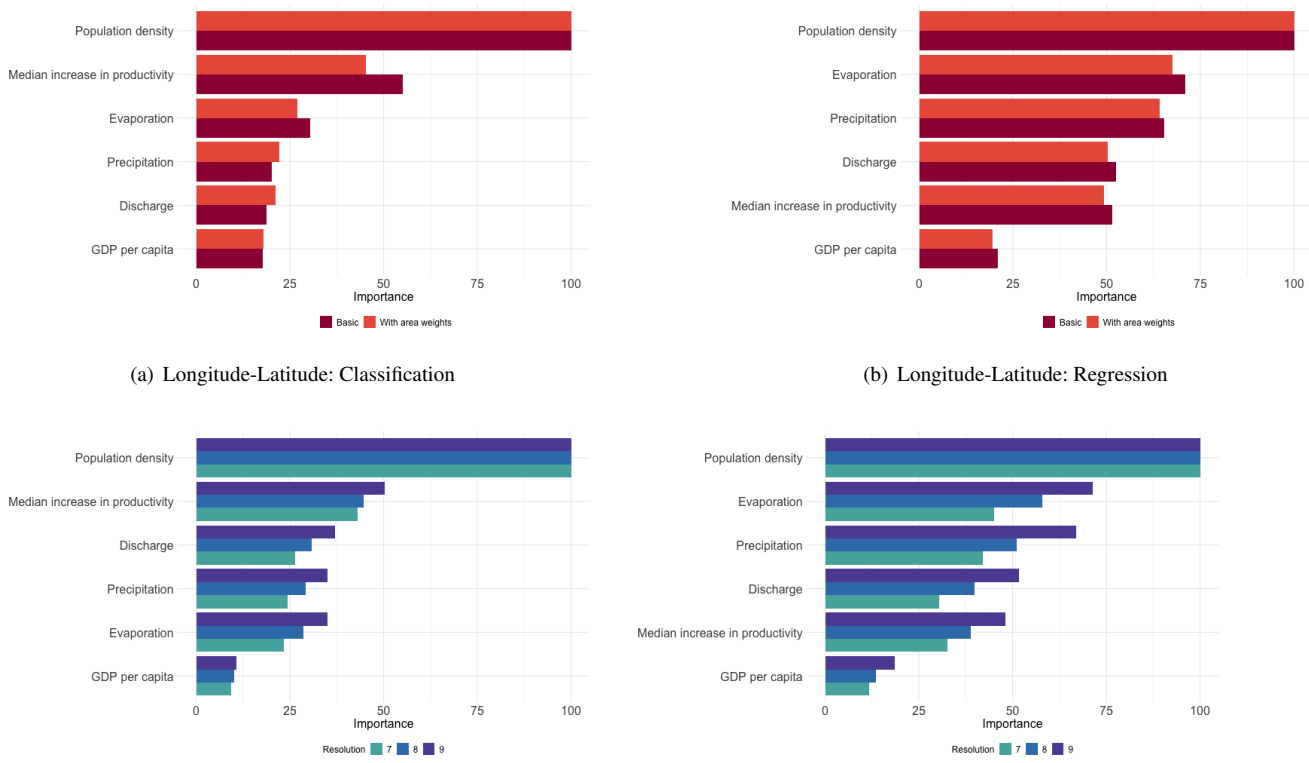
**Figure 3.** Deviation of the predicted irrigation fraction from the observed irrigation fraction in (a) the longitude-latitude grid and (b)-(d)the ISEA3H grids at resolutions 7, 8, and 9. Green and blue areas indicate under-prediction of the irrigation value and orange and red values over-prediction. Yellow areas correspond to areas where irrigation values were predicted correctly. The prediction is based on the test data.

### 4.2 Drivers of irrigation expansion

#### 4.2.1 Variable importance

We report the importance of the predictors for both the classification random forests, which predict the probability of irrigation occurring, and the regression random forests, which predict irrigation magnitude given that the area is irrigated. In the classification random forests, relative importance is measured by Gini gain, while in the regression random forests, it is captured by the estimated response variance. The results are displayed in Figure 4.

The most important driver for the probability that an area is irrigated is population density. This is the case for all grid choices.




(a) Longitude-Latitude: Classification

(b) Longitude-Latitude: Regression

(c) ISEA3H: Classification

(d) ISEA3H: Regression

**Figure 4.** Relative importance, measured as decrease of node impurity. The results for the longitude-latitude grids can be seen in red and the results for the ISEA3H grids are displayed in blue. The order of variables in the importance plots are robust to 500 bootstrap steps.

The second most important driver is the median potential increase in productivity in terms of crop yield. Evaporation, precipitation and discharge all have a similar influence on irrigation probability. However, the order of importance is reversed between the two grid choices. GDP per capita only has small influence on the decision to irrigate.

The most important driver of irrigation intensity, given that an area is already irrigated, is also population density. This is followed by evaporation, precipitation, discharge and the median increase in potential productivity, where the order of discharge

and the median potential productivity increase is reversed for the ISEA3H grids. The last most important driver is again GDP per capita, though still having some influence on the models' performance in all grids.

Looking at the different patterns across resolutions, it appears that finer resolutions increase the relative influence of predictors other than population density. This makes sense, as these other drivers likely have a greater impact at a local level, whereas population density reflects a broader need for crop production in the area.

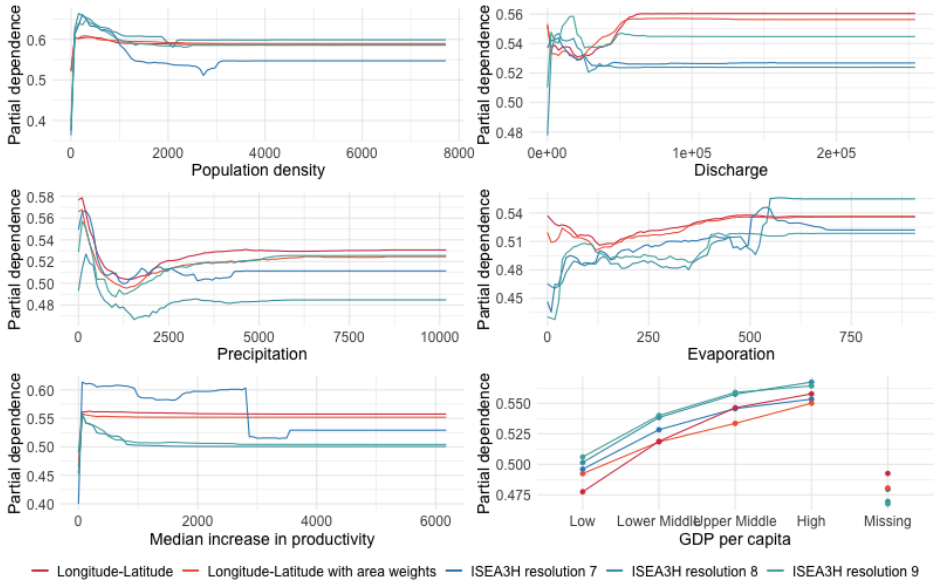
#### 4.2.2 Partial dependence

We compute the partial dependence of each predictor variable for all grid choices and model specifications. The partial dependence is obtained by gradually changing the value of one predictor variable and predicting the outcome variable at each step, while leaving the remaining predictors constant. That way, the functional relationship between the predictor and the dependent variable becomes visible. The larger the value range on the vertical axis, the larger the influence of the predictor on the dependent variable. Figure 5 illustrates the results.

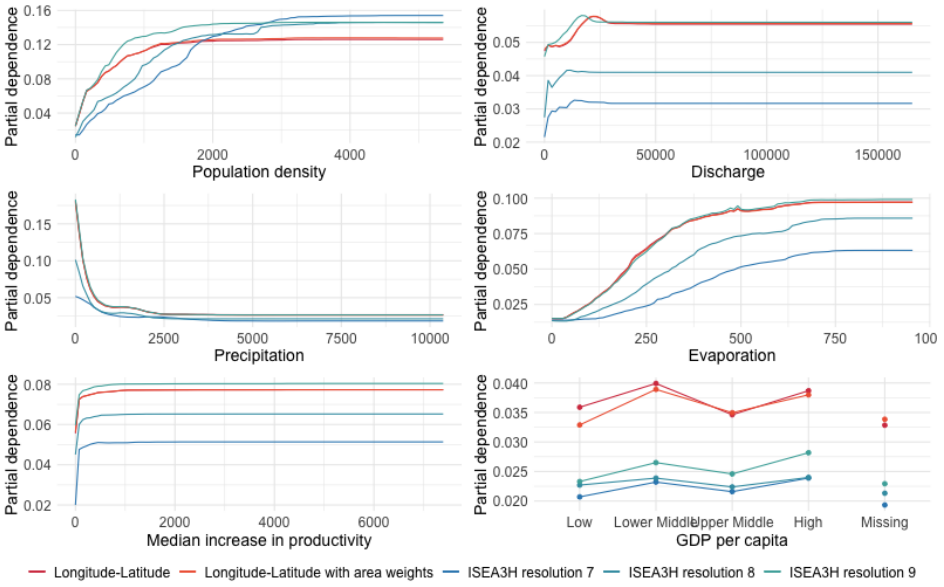
Panel a) of Figure 5 illustrates the partial dependence of the predictors of irrigation probability. Overall, we see very intuitive dependence patterns. Population density has a positive influence on the probability to irrigate, where the probability sharply increases at the beginning of the population density distribution. In other words, greater population density correlates with an increased probability of irrigation, indicating that metropolitan regions with higher population densities and improved market accessibility are more likely to engage in irrigation. This heightened probability is likely attributed to the requirement of capital investment for establishing irrigation systems. This aligns with the paper by Neumann et al. (2011), who also found a positive association between irrigation and population density.

A similar pattern can be seen for the median potential increase in productivity, the second most influential predictor. The positive correlation demonstrates that the potential increase in crop yield is a factor for the decision to implement irrigation systems.

Evaporation also has a positive, almost linearly increasing influence on irrigation probability. Considering precipitation, our results show that the probability to irrigate decreases with the amount of precipitation until the probability levels off. The amount of available discharge has a negative relationship with the probability to irrigate for both grid choices at the beginning of the distribution. Looking at the longitude-latitude grid, this changes into a positive correlation, leaving us with a u-shaped dependence curve. Looking at the ISEA3H grid choice, the irrigation probability does not change anymore after reaching a certain discharge level. Overall, these results show that water availability and climatic conditions play a role for the decision to irrigate, leaving rather dry areas and areas with higher evaporation levels more likely to be or become irrigated. Discharge is an accumulated variable of local runoff, with very high differences between upstream and downstream cells in a watershed. This means that regions with relatively high topography and thus potentially lower degrees of agriculture and irrigation are all coinciding with low discharge values, while the major irrigation areas (India, Pakistan, US, East-Asia, Egypt, ...) generally lie close to large streams with high discharge. The correlation with elevation might explain why initially the dependence of irrigation on discharge decreases. For large values the large grid-size might be able to explain the differences between the grid, as for example along the Nile the irrigated areas follow the river in a small band, being dispersed in the ISEA3H grid.




(a) Classification




(b) Regression

**Figure 5.** Partial dependence of the predictors and the dependent variable of (a) the classification random forests and (b) the regression random forests. The results for the ISEA3H grids (resolution 7, 8, and 9) are shown in shades of blue and the result for the longitude-latitude grid and the model with area weights are shown in red.

Lastly, we study the dependence of the GDP per capita categories on the probability to observe irrigation. We find a strictly positive relationship from the categories ”Low”, ”Lower Middle”, ”Upper Middle” to ”High”. Therefore, the likelihood of croplands being irrigated is higher for areas with generally higher economic performance. Hence, adverse socio-economic conditions hinder the development of irrigated agriculture. This result complements the findings of Neumann et al. (2011), who found similar effects considering government performance and government type. The GDP per capita category ”missing” corresponds to a relatively lower irrigation probability. This is in line with the fact that in earlier time periods, less areas were irrigated and more GDP per capita observations are missing.

Panel b) of Figure 5 displays the partial dependence curves for the predictor variables of irrigation intensity, i.e. the amount of irrigation given a grid cell is irrigated. The most influential predictor, population density, positively impacts the amount of irrigation.

Evaporation is also positively associated with the amount of irrigation, where the increase in irrigation appears to be almost linear in evaporation levels. The amount of irrigation negatively depends on precipitation levels, while discharge is positively correlated with irrigation intensity. Hence, the effect of water availability differs between different sources of water, where heavily precipitated areas do not seem to require as much irrigation, while discharge might be used to feed irrigation systems.

The median potential productivity gain is positively associated with irrigation intensity, exhibiting a sharp peak in the dependence curve at the beginning of the distribution. Much of the tails is probably irrelevant for a real-world scenario, where irrigation would never happen in remote and dry regions, with a high potential for productivity increases from irrigation. Larger cell sizes in the ISEA3H grid mean "easier" access to streams (more area is in the same cell as the river), which is reflected in the higher plateau level.

Considering GDP per capita, we see irrigation intensity only slightly differing between the categories.

Assessing our results in the context of our hypotheses (see Table 1), we generally observe a consistent alignment between our empirical results and our previous theoretical consideration.

## 5 Conclusions

The careful choice of a discrete global grid system holds significant importance for conducting statistical analyses on a global scale. In this paper, we make use of historical global irrigation data from the last century, to compare the standard longitude-latitude grid to ISEA3H discrete global grids at different resolutions. We employ a stacked random forest framework to model probability of irrigation and irrigation magnitude (once an area is irrigated) as a function of potential drivers. We identify population density and the potential productivity increase in terms of crop yield as the most influential factors for the decision to irrigate and population density and factors accounting for water availability as drivers for intense irrigation. We further point to GDP per capita as having some influence on irrigation behaviour.

Comparing the two grid systems, we find that ISEA3H geodesic discrete global grids yield higher prediction accuracies. Using the assigned test data, the model built on the geodesic discrete global grid at resolution 7 produces a 28% lower root normalised mean squared prediction error compared to the model built on the longitude-latitude grid. Although the difference

in predictive accuracy decreases with higher resolutions, the ISEA3H grids at resolutions 8 and 9 still produce significantly lower error values compared to the benchmark model. In comparison, using traditional area weights in the longitude-latitude grid does improve prediction accuracy significantly. These results are robust to different normalisation definitions.

In terms of the global irrigation prediction pattern, we find that the models based on the ISEA3H grids come closer to the observed irrigation map. While the longitude-latitude grid leads to some highly under-predicted areas in India, East-Asia and the United States, the ISEA3H grids are associated with under-prediction in almost the same areas, although smaller in magnitude. Although the increase in predictive accuracy might partly be due to the fact that the change in grid cell structure changes the scale and therefore the range of values of the targeted irrigation variable, the advantages of the uniformly structured ISEA3H grids are evident and should be explored and tested in future research.

While the combination of water availability, climate, and socioeconomic data offer valuable insights into the role of discrete global grid choice and the drivers of historical irrigation expansion, it is clear that our setting does not come without limitations. For example, we neglected seasonality, meaning that yearly values were used for the analysis. However in reality water availability is much more relevant in the growing season than in the off season. While we offer new evidence about the potential accuracy increase using a geodesic discrete global grid, our methodology does not include an exhaustive search for the best-possible grid choice. Our goal is rather to set a first reference point for future research designs.

We model irrigation fraction as a function of precipitation, discharge, evaporation, population density, potential productivity increase in terms of crop yield, and GDP per capita. While these are important drivers of irrigation, there are likely other contributing factors that we are not able to capture in our analysis, such as the access to groundwater, irrigation subsidies, or other socioeconomic factors such as the type of government. The access to spatially explicit information would allow researchers to further explore these potential drivers.

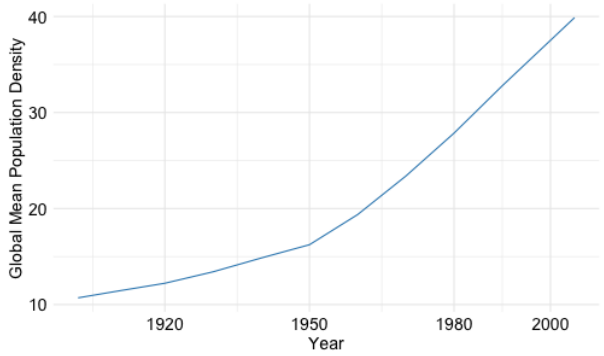
Another interesting avenue for future research is to include time-lags in the analysis. It might not be the data of the same year (e.g. 1990) that are most indicative of the irrigation fraction of that year, but for example the (average) data of the previous decade. These time-lags might even be different for different predictors.

Lastly, the irrigation and predictor data are based on a large variety of sources from different years, which have likely introduced uncertainties.

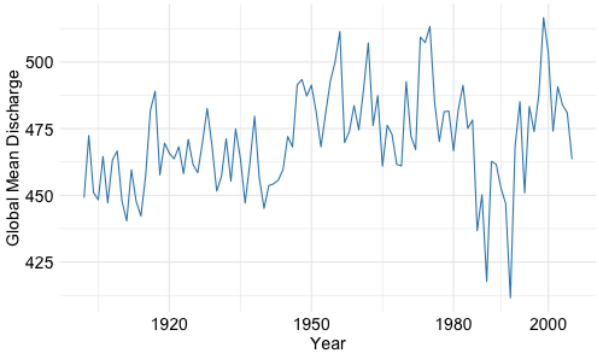
Acknowledging these limitations, we consider our analysis as an important step towards understanding the role of discrete global grids in global statistical modelling. Particularly, exploring the application of the ISEA3H geodesic grid system in different global analytical contexts presents an intriguing avenue for future research.

*Code and data availability.* The code and data used in this study are publicly available for download at Zenodo https://doi.org/10.5281/zenodo.12542249.

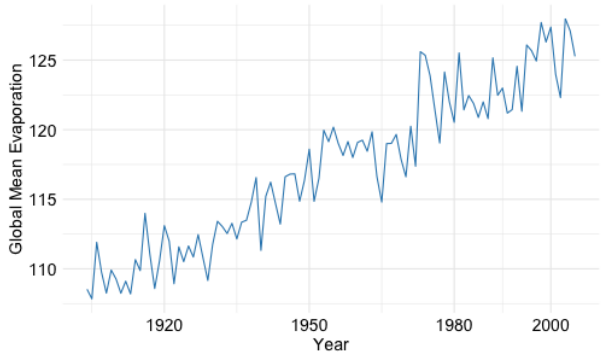
## Appendix A



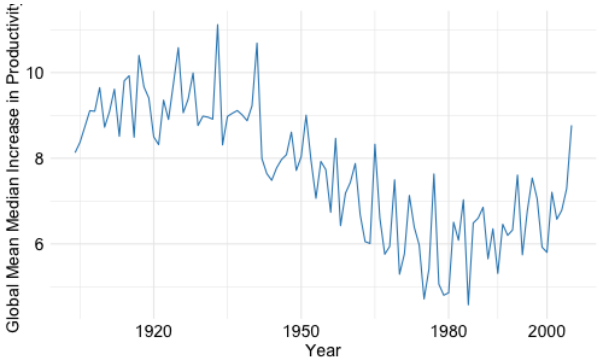

(a) Population density



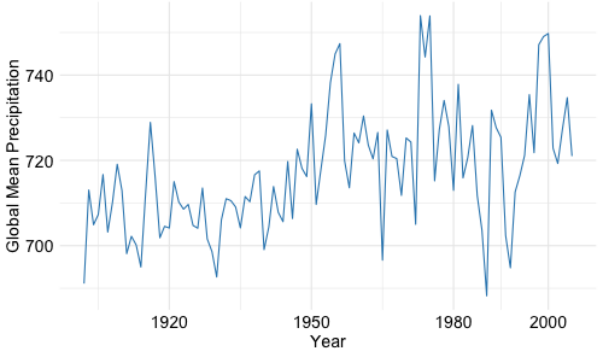

(b) Discharge



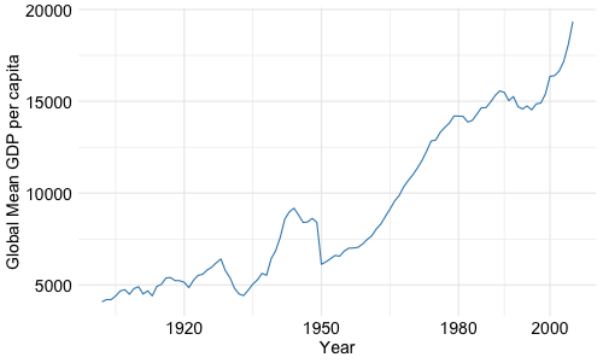

(c) Evaporation




(d) Median increase in productivity

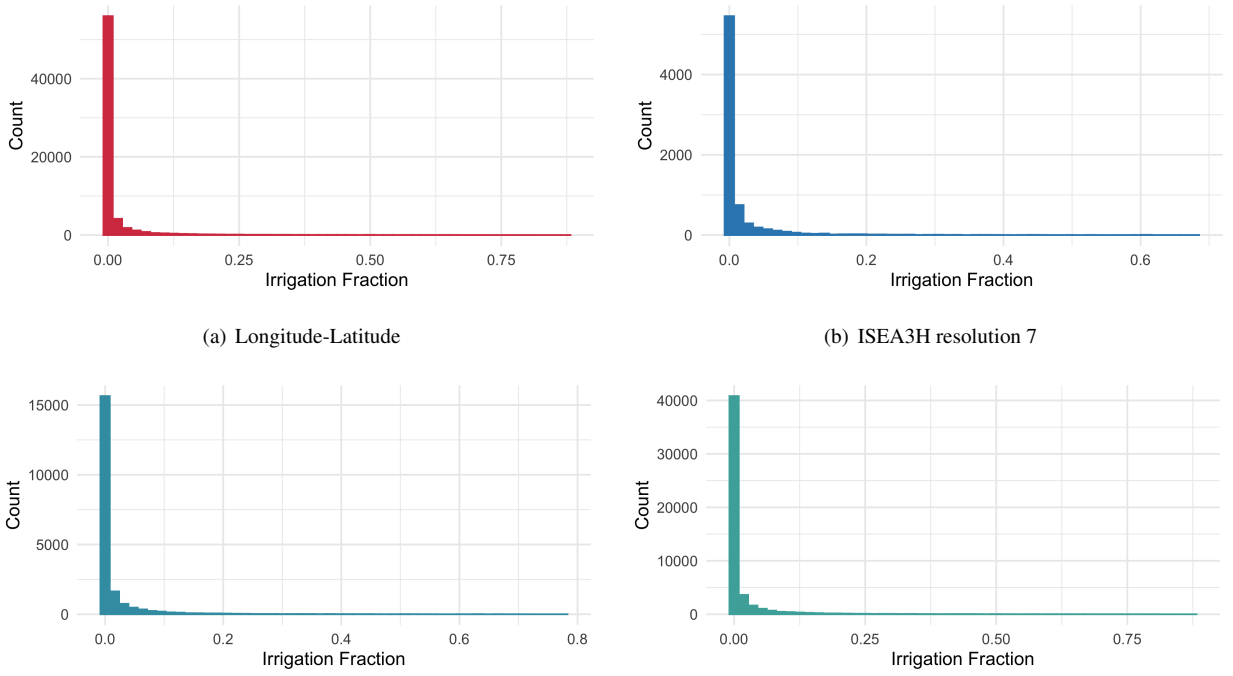



(e) Precipitation




(f) GDP per capita

**Figure A1.** Evolution of the global means of the predictor variables across the study period 1902 to 2005.




(a) Longitude-Latitude

(b) ISEA3H resolution 7

(c) ISEA3H resolution 8

(d) ISEA3H resolution 9

**Figure A2.** Histograms, showing the irrigation fraction on the horizontal axes and the corresponding frequency of the observational data used in the analysis in (a) the longitude-latitude grid and (b)-(d) the ISEA3H grids (at resolutions 7, 8, and 9) on the vertical axes.

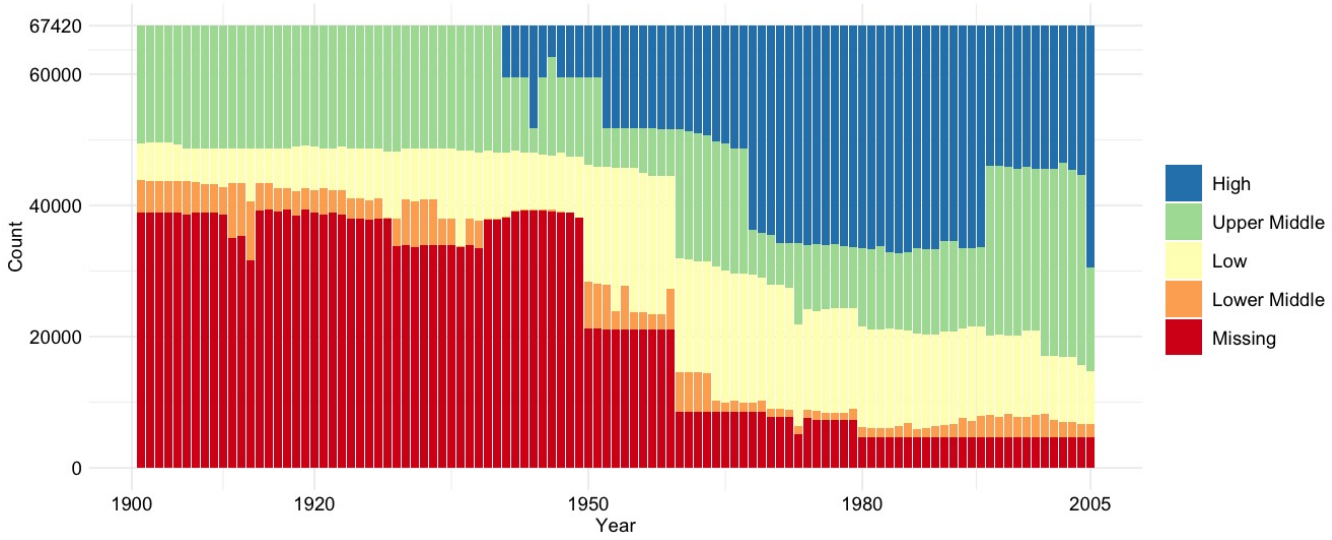
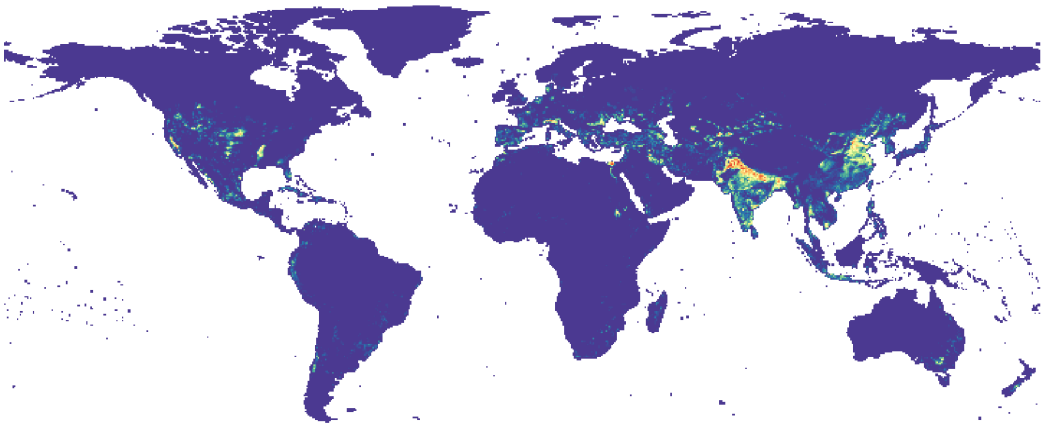
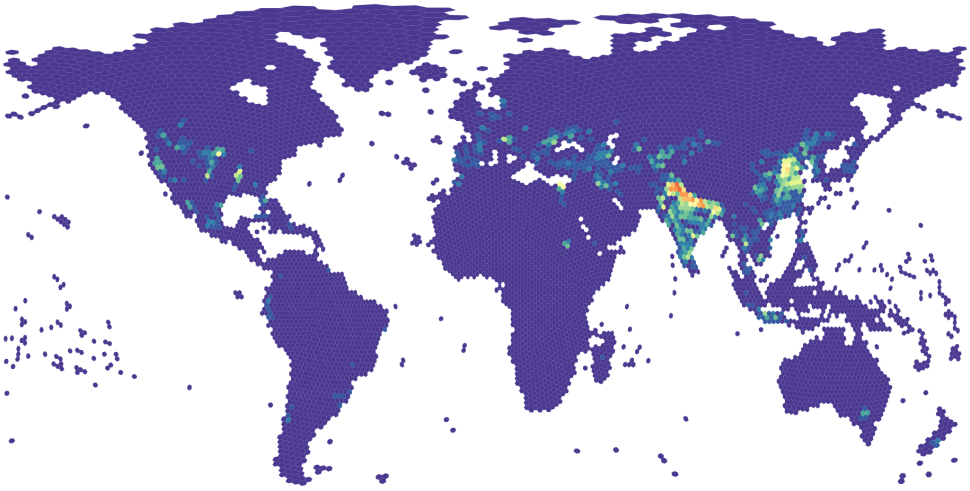


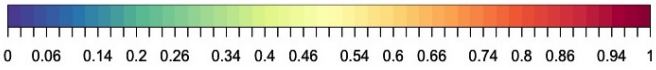

**Figure A3.** Frequency of GDP per capita categories over the study period 1902 to 2005.



(a) Longitude-Latitude



(b) ISEA3H resolution 7




(c)

**Figure A4.** Irrigation fraction in 2000 in a) the longitude-latitude discrete global grid and b) the ISEA3H (resolution 7) discrete global grid. Irrigation fraction reflects the area irrigated of each grid cell and is based on the global Historical Irrigation Dataset (see Siebert et al., 2015).

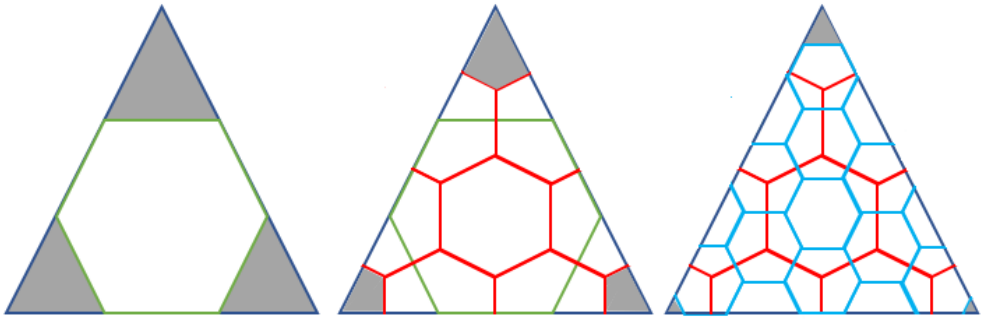
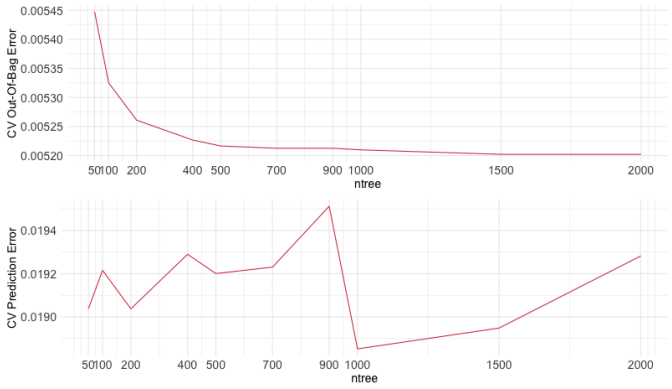
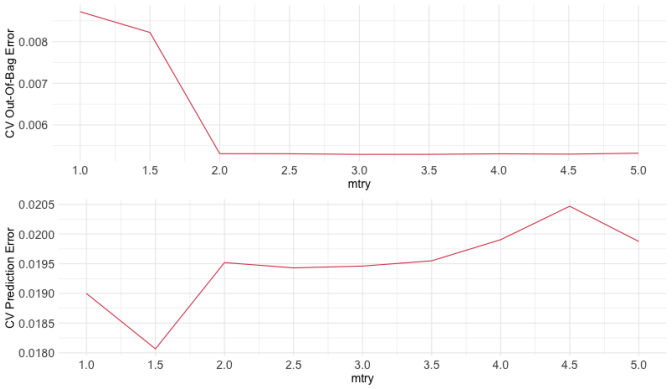


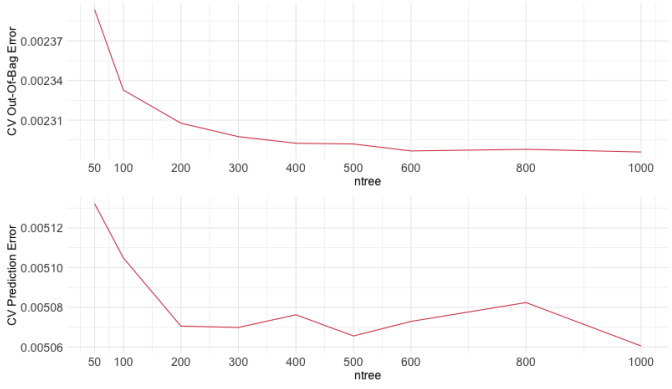
**Figure A5.** Recursive partitioning aperture 3 method. The hexagonal pattern is recursively constructed on top of the base icosahedron. The first resolution is illustrated by the green hexagon, directly constructed inside a triangular face of the base icosahedron. The construction of the resolution 2 grid is displayed in red in the middle. The resolution 3 hexagonal pattern is illustrated on the right side. Increasing the resolution by one, leads to hexagons with a size of one third of the original hexagon size. The grey left over areas are the reason why overall, a few pentagonal faces are needed to cover the Earth's surface. The image is based on an illustration by de Wiljes (2015).



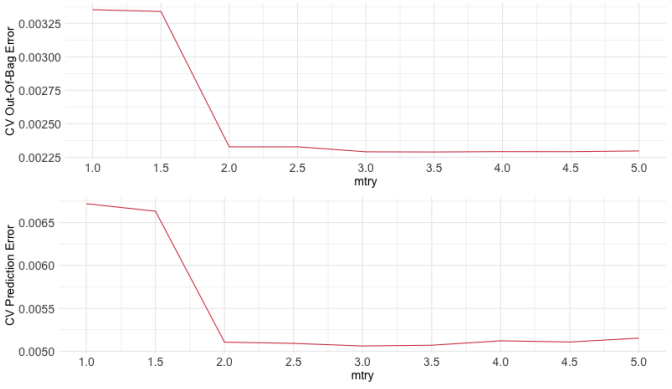

(a) Longitude-Latitude Classification: Ntree




(b) Longitude-Latitude Classification: Mtry

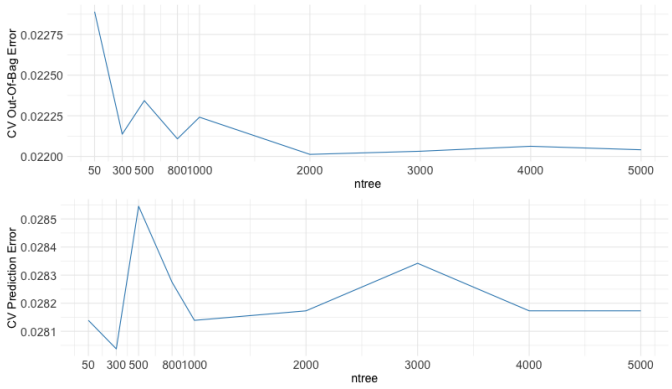



(c) Longitude-Latitude Regression: Ntree

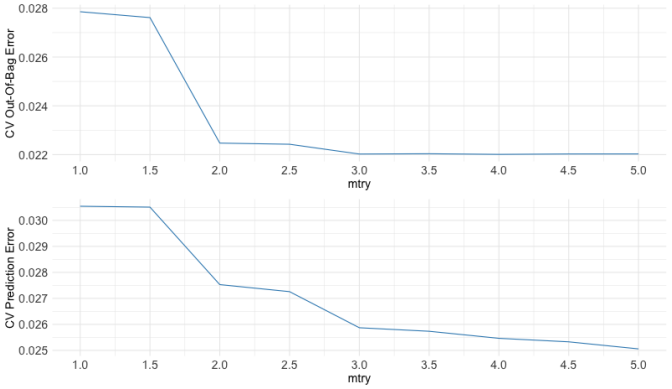



(d) Longitude-Latitude Regression: Mtry

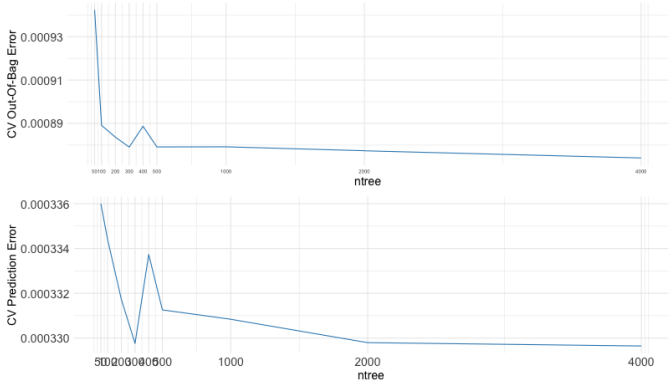
**Figure A6.** Cross-validation results of the longitude-latitude grid choice. The out-of-bags error and the prediction error are displayed as a function of changing hyperparameter values for a) ntree in the classification random forest, b) mtry in the classification random forest, c) ntree in the regression random forest and d) mtry in the regression random forest.



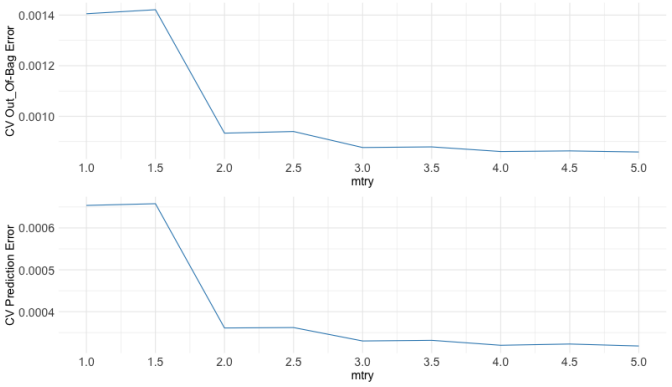

(a) ISEA3H Classification: Ntree




(b) ISEA3H Classification: Mtry




(c) ISEA3H Regression: Ntree




(d) ISEA3H Regression: Mtry

**Figure A7.** Cross-validation results of the geodesic discrete global grid choice. The out-of-bags error and the prediction error are displayed as a function of changing hyperparameter values for a) ntree in the classification random forest, b) mtry in the classification random forest, c) ntree in the regression random forest and d) mtry in the regression random forest.

**Table A1.** Summary statistics of the training data (1902-1999)

| | Mean | Standard deviation | Minimum | Maximum | Median |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| **A. Longitude-Latitude grid (n = 6.607,160)** | | | | | |
| Irrigation fraction (%) | 0.0077 | 0.0375 | ~~0.0000~~ 0 | 0.9220 | ~~0.00000~~ |
| Population density (cap/$m^2$) | 19.5986 | 72.4894 | ~~0.0000~~ 0 | ~~9832.0000~~ 9832 | ~~1.00001~~ |
| Precipitation (mm/year) | 716.3860 | 712.2138 | ~~0.0000~~ 0 | ~~11155.0000~~ 11155 | 478.9372 |
| Evaporation (mm/year) | 116.6513 | 80.7215 | ~~0.0000~~ 0 | 953.9896 | 97.3343 |
| Discharge ($hm^2$/year) | 469.6246 | 4524.4351 | ~~0.00000~~ | 270078.8232 | 28.0981 |
| Median increase in productivity (% of Δ $gC/m^2$) | 7.6580 | 65.9509 | -0.5596 | 17365.5508 | 0.0053 |
| **B. ISEA3H grid res. 7 (n = 730,917)** | | | | | |
| Irrigation fraction (%) | 0.0084 | 0.0318 | ~~0.0000~~ 0 | 0.8077 | ~~0.00000~~ |
| Population density (cap/$m^2$) | 23.5708 | 69.3340 | ~~0.0000~~ 0 | ~~4575.0000~~ 4575 | ~~2.00002~~ |
| Precipitation (mm/year) | 905.5131 | 848.5934 | ~~0.0000~~ 0 | ~~10853.0000~~10853 | 609.2273 |
| Evaporation (mm/year) | 134.0022 | 84.9666 | ~~0.0000~~ 0 | 715.0584 | 116.7027 |
| Discharge ($hm^2$/year) | 517.9301 | 3245.9758 | ~~0.00000~~ | 134255.6759 | 70.7079 |
| Median increase in productivity (% of Δ $gC/m^2$) | 8.9070 | 53.7531 | -0.1054 | 4313.1124 | 0.0421 |

Notes: Panel A summarizes the descriptive statistics of the test data set in the original longitude-latitude grid. The test data set contains the years 1902 to 1999. Panel B summarizes the descriptive statistics of the ISEA3H grid, i.e. after transforming the data to the hexagonal grid. The GDP per capita predictor is excluded from this summary table, as it is a factor variable.

**Table A2.** GDP per capita category assignment

| Class | GDP Per Capita |
|---|---|
| High | $\geq$ 12,276$ |
| Upper Middle | > 3,975$ – 12,275$ |
| Lower Middle | > 1,005$ – 3,975$ |
| Low | $\leq$ 1,005$ |
| Missing | – |

Notes: GDP per capita classification by income level for the reference year 2011, based on the classification of the World Bank (2011).

**Table A3.** Pearson correlation coefficient

| | Pearson Correlation Coefficient | | | |
|---|---|---|---|---|
| | Population density | Median increase in productivity | Discharge | Precipitation |
| | (1) | (2) | (3) | (4) |
| Median increase in productivity | −0.0212 | | | |
| Discharge | 0.0116 | −0.009 | | |
| Precipitation | 0.1349 | −0.094 | 0.1111 | |
| Evaporation | 0.2403 | −0.0417 | 0.0588 | 0.720 |

Notes: In this table, the correlation matrix of the Pearson correlation coefficient of the predictors is presented. The displayed values are the lower half of the correlation matrix.

**Table A4.** Variance inflation factor

| | Variance Inflation Factor | | | | |
|---|---|---|---|---|---|
| | Population density | Median increase in productivity | Discharge | Precipitation | Evaporation |
| | (1) | (2) | (3) | (4) | (5) |
| VIF | 1.065230 | 1.010816 | 1.013481 | 2.124101 | 2.183165 |

Notes: This table displays the variance inflation factor (VIF) of the predictor variables. The measure is used to detect multicollinearity between potential predictor variables. A VIF below 5, means that the respective variable is not collinear to the other variables (James et al. (2013)).

*Author contributions.* SW, FS, TK and JdW conceptualized the underlying research and designed a choice of models and methodologies for data analysis. SW performed the research and led the project. SW, FS, TK and JdW analyzed the results. SW wrote the original draft of the manuscript. FS, TK and JdW reviewed and edited the manuscript.

*Competing interests.* The contact author has declared that none of the authors has any competing interests.

*Acknowledgements.* This research has been partially funded by Deutsche Forschungsgemeinschaft (DFG) - SFB1294/1 - 318763901. FS acknowledges funding by the CE-Land+ project of the German Research Foundation's priority program SPP 1689 "Climate Engineering – Risks, Challenges and Opportunities?" as well as by the Global Challenges Foundation via Future Earth.

## References


Apte, M., Y.y., A. Y., S., A., and B., I. A.: Article: Understanding Grids and Effectiveness of Hexagonal Grid in Spatial Domain, IJCA Proceedings on International Conference on Recent Trends in Information Technology and Computer Science 2012, pp. 25–27, full text available, 2013.

Barnes, R. and Sahr, K.: dggridR: Discrete Global Grids for R, R package version 2.0.4., https://doi.org/doi:10.5281/zenodo.1322866, 2017.

Bernard, S., Heutte, L., and Adam, S.: Influence of Hyperparameters on Random Forest Accuracy, in: MCS, 2009.

Bolt, J. and Zanden, J. L.: The Maddison Project: collaborative research on historical national accounts, Economic History Review, 67, 627–651, https://EconPapers.repec.org/RePEc:bla:ehsrev:v:67:y:2014:i:3:p:627-651, 2014.

Bondaruk, B., Roberts, S. A., and Robertson, C.: Assessing the state of the art in Discrete Global Grid Systems: OGC criteria and present functionality, Geomatica, 74, 9–30, https://doi.org/10.1139/geomat-2019-0015, 2020.

Bondeau, A., Smith, P. C., Zaehle, S., Schaphoff, S., Wolfgang, L., Cramer, W., Gerten, D., Lotze-Campen, H., Müller, C., Reichstein, M., and Smith, B.: Modelling the role of agriculture for the 20th century global terrestrial carbon balance, Global Change Biology, 13, 679–706, https://doi.org/https://doi.org/10.1111/j.1365-2486.2006.01305.x, 2007.

Boretti, A. and Rosa, L.: Reassessing the projections of the World Water Development Report, npj Clean Water, 2, https://doi.org/10.1038/s41545-019-0039-9, 2019.

Bousquin, J.: Discrete Global Grid Systems as scalable geospatial frameworks for characterizing coastal environments, Environmental Modelling & Software, 146, 105 210, https://doi.org/https://doi.org/10.1016/j.envsoft.2021.105210, 2021.

Breiman, L.: Random Forests, Mach. Learn., 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

Budic, L. et al.: Squares of different sizes: effect of geographical projection on model parameter estimates in species distribution modeling, Ecology and Evolution, 6, 202–211, https://doi.org/https://doi.org/10.1002/ece3.1838, 2016.

Chaudhuri, C., Gray, A., and Robertson, C.: InundatEd-v1.0: a height above nearest drainage (HAND)-based flood risk modeling system using a discrete global grid system, Geoscientific Model Development, 14, 3295–3315, https://doi.org/10.5194/gmd-14-3295-2021, 2021.

de Wiljes, J.: Data-Driven Discrete Spatio-Temporal Models: Problems, Methods and an Arctic Sea Ice Application, Ph.D. thesis, Humboldt University Berlin, http://dx.doi.org/10.17169/refubium-4258, 2015.

Döll, P. and Siebert, S.: Global modeling of irrigation water requirements, Water Resources Research, 38, 8–1–8–10, https://doi.org/https://doi.org/10.1029/2001WR000355, 2002.

Döll, P., Mueller-Schmied, H., Schuh, C., Portmann, F. T., and Eicker, A.: Global-scale assessment of groundwater depletion and related groundwater abstractions: Combining hydrological modeling with information from well observations and GRACE satellites, Water Resources Research, 50, 5698–5720, https://doi.org/https://doi.org/10.1002/2014WR015595, 2014.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., and Lautenbach, S.: Collinearity: a review of methods to deal with it and a simulation study evaluating their performance, Ecography, 36, 27–46, https://doi.org/https://doi.org/10.1111/j.1600-0587.2012.07348.x, 2013.

FAO, F. and of the United Nations, A. O.: State of the Worlds Land and Water Resources for Food and Agriculture: Managing systems at risk. Food and Agriculture Organization of the United Nations, Food & Agriculture Org, 2011.

Foley, J., Ramankutty, N., Brauman, K., Cassidy, E., Gerber, J., Johnston, M., Mueller, N., O'Connell, C., Ray, D., West, P., Balzer, C., Bennett, E., Carpenter, S., Hill, J., Monfreda, C., Polasky, S., Rockström, J., Sheehan, J., Siebert, S., and Zaks, D.: Solutions for a Cultivated Planet, Nature, 478, 337–342, https://doi.org/10.1038/nature10452, 2011.

Gerten, D., Rost, S., von Bloh, W., and Lucht, W.: Causes of change in 20th century global river discharge, Geophysical Research Letters, 35, https://doi.org/https://doi.org/10.1029/2008GL035258, 2008.

Goodchild, M. F.: Geographical grid models for environmental monitoring and analysis across the globe (panel session), in: Proceedings of GIS/LIS. ACSM & ASPRS, Bethesda, Maryland, 1994.

Hahmann, S. and Burghardt, D.: How much information is geospatially referenced? Networks and cognition, International Journal of Geographical Information Science, 27, 1171–1189, https://doi.org/10.1080/13658816.2012.743664, 2013.

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

Hastie, T., Tibshirani, R., and Friedman, J.: The elements of statistical learning: data mining, inference and prediction, Springer, http://www-stat.stanford.edu/~tibs/ElemStatLearn/, 2009.

Hojati, M. and Robertson, C.: Integrating cellular automata and discrete global grid systems: a case study into wildfire modelling, AGILE: GIScience Series, 1, 6, https://doi.org/10.5194/agile-giss-1-6-2020, 2020.

Huang, B. F. and Boutros, P. C.: The parameter sensitivity of random forests, BMC Bioinformatics, 17, https://doi.org/10.1186/s12859-016-1228-x, 2016.

Inklaar, R., de Jong, H., Bolt, J., and van Zanden, J.: Rebasing 'Maddison': new income comparisons and the shape of long-run economic development, Ggdc research memorandum, Groningen Growth and Development Center, https://EconPapers.repec.org/RePEc:gro:rugggd:gd-174, 2018.

ISO: ISO 19170-1:2021, https://www.iso.org/standard/32588.html, 2021.

James, G., Witten, D., Hastie, T., and Tibshirani, R.: An Introduction to Statistical Learning: with Applications in R, Springer New York, https://faculty.marshall.usc.edu/gareth-james/ISL/, 2013.

Jendryke, M. and McClure, S. C.: Mapping crime - Hate crimes and hate groups in the USA: A spatial analysis with gridded data, Applied Geography, 111, 102 072, https://doi.org/https://doi.org/10.1016/j.apgeog.2019.102072, 2019.

Jägermeyr, J., Gerten, D., Heinke, J., Schaphoff, S., Kummu, M., and Lucht, W.: Water savings potentials of irrigation systems: global simulation of processes and linkages, Hydrology and Earth System Sciences, 19, 3073–3091, https://doi.org/10.5194/hess-19-3073-2015, 2015.

Klein Goldewijk, K., Beusen, A., Doelman, J., and Stehfest, E.: Anthropogenic land use estimates for the Holocene – HYDE 3.2, Earth System Science Data, 9, 927–953, https://doi.org/10.5194/essd-9-927-2017, 2017.

Li, M., McGrath, H., and Stefanakis, E.: Integration of heterogeneous terrain data into Discrete Global Grid Systems, Cartography and Geographic Information Science, 48, 546–564, https://doi.org/10.1080/15230406.2021.1966648, 2021.

Li, M., McGrath, H., and Stefanakis, E.: Geovisualization of Hydrological Flow in Hexagonal Grid Systems, Geographies, 2, 227–244, https://doi.org/10.3390/geographies2020016, 2022.

Liao, C., Tesfa, T. K., Duan, Z., and Leung, L.-Y.: Watershed Delineation On A Hexagonal Mesh Grid, Environmental Modelling & Software, 128, https://doi.org/10.1016/j.envsoft.2020.104702, 2020.

Lunardon, N. et al.: ROSE: a Package for Binary Imbalanced Learning, R Journal, 6, 82–92, 2014.

McPhail, C. K.: RECONSTRUCTING ERATOSTHENES' MAP OF THE WORLD: A STUDY IN SOURCE ANALYSIS, Ph.D. thesis, University of Otago, 2011.

Mechenich, M. and Zliobaite, I.: Eco-ISEA3H, a machine learning ready spatial database for ecometric and species distribution modeling, Scientific data, 10, https://doi.org/10.1038/s41597-023-01966-x, 2023.

Meier, J., Zabel, F., and Mauser, W.: A global approach to estimate irrigated areas – a comparison between different data and statistics, Hydrology and Earth System Sciences, 22, 1119–1133, https://doi.org/10.5194/hess-22-1119-2018, 2018.

Nagaraj, D., Proust, E., Todeschini, A., Rulli, M. C., and D'Odorico, P.: A new dataset of global irrigation areas from 2001 to 2015, Advances in Water Resources, 152, 103 910, https://doi.org/https://doi.org/10.1016/j.advwatres.2021.103910, 2021.

Neumann, K., Stehfest, E., Verburg, P. H., Siebert, S., Müller, C., and Veldkamp, T.: Exploring global irrigation patterns: A multilevel modelling approach, Agricultural Systems, 104, 703–713, https://doi.org/https://doi.org/10.1016/j.agsy.2011.08.004, 2011.

Oshiro, T. M., Perez, P. S., and Baranauskas, J. A.: How Many Trees in a Random Forest?, in: Machine Learning and Data Mining in Pattern Recognition, edited by Perner, P., pp. 154–168, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

Probst, P., Wright, M. N., and Boulesteix, A.-L.: Hyperparameters and tuning strategies for random forest, WIREs Data Mining and Knowledge Discovery, 9, e1301, https://doi.org/https://doi.org/10.1002/widm.1301, 2019.

Purss, M.: Topic 21: Discrete global grid systems abstract specification, http://docs.opengeospatial.org/as/15-104r5/15-104r5.html, 2015.

Puy, A., Lo Piano, S., and Saltelli, A.: Current Models Underestimate Future Irrigated Areas, Geophysical Research Letters, 47, https://doi.org/https://doi.org/10.1029/2020GL087360, 2020.

Robertson, C., Chaudhuri, C., Hojati, M., and Roberts, S. A.: An integrated environmental analytics system (IDEAS) based on a DGGS, ISPRS Journal of Photogrammetry and Remote Sensing, 162, 214–228, https://doi.org/https://doi.org/10.1016/j.isprsjprs.2020.02.009, 2020.

Rufin, P., Levers, C., Baumann, M., Jägermeyr, J., Krueger, T., Kuemmerle, T., and Hostert, P.: Global-scale patterns and determinants of cropping frequency in irrigation dam command areas, Global Environmental Change, 50, 110–122, https://doi.org/https://doi.org/10.1016/j.gloenvcha.2018.02.011, 2018.

Sahr, K.: Hexagonal discrete global grid systems for geospatial computing, Archives of Photogrammetry, Cartography and Remote Sensing, 22, 2011.

Sahr, K. et al.: Geodesic Discrete Global Grid Systems, Cartography and Geographic Information Science, 30, 121–134, https://doi.org/10.1559/152304003100011090, 2003.

Sahr, K. et al.: The PlanetRisk Discrete Global Grid System, Department of Computer Science, Southern Oregon University, https://www.discreteglobalgrids.org/publications/, 2015.

Salmon, J. M., Friedl, M. A., Frolking, S., Wisser, D., and Douglas, E. M.: Global rain-fed, irrigated, and paddy croplands: A new high resolution map derived from remote sensing, crop inventories and climate data, Int. J. Appl. Earth Obs. Geoinformation, 38, 321–334, 2015.

Sauer, T., Havlik, P., Schneider, U. A., Schmid, E., Kindermann, G., and Obersteiner, M.: Agriculture and resource availability in a changing world: The role of irrigation, Water Resources Research, 46, https://doi.org/https://doi.org/10.1029/2009WR007729, 2010.

Schaphoff, S., von Bloh, W., Rammig, A., Thonicke, K., Biemans, H., Forkel, M., Gerten, D., Heinke, J., Jägermeyr, J., Knauer, J., Langerwisch, F., Lucht, W., Mueller, C., Rolinski, S., and Waha, K.: LPJmL4 – a dynamic global vegetation model with managed land – Part 1: Model description, Geoscientific Model Development, 11, 1343–1375, https://doi.org/10.5194/gmd-11-1343-2018, 2018.

Schultz, B. et al.: Irrigation and drainage. Main contributors to global food production, Irrigation and Drainage, 54, 263–278, https://doi.org/https://doi.org/10.1002/ird.170, 2005.

Siebert, S., Kummu, M., Porkka, M., Doell, P., Ramankutty, N., and Scanlon, B. R.: A global data set of the extent of irrigated land from 1900 to 2005, Hydrology and Earth System Sciences, 19, 1521–1545, https://doi.org/10.5194/hess-19-1521-2015, 2015.

Sirdeshmukh, N., Verbree, E., Oosterom, P. V., Psomadaki, S., and Kodde, M.: Utilizing a Discrete Global Grid System for Handling Point Clouds with Varying Locations, Times, and Levels of Detail, Cartographica: The International Journal for Geographic Information and Geovisualization, 54, 4–15, https://doi.org/10.3138/cart.54.1.2018-0009, 2019.

Sitch, S., Smith, B., Prentice, I. C., Arneth, A., Bondeau, A., Cramer, W., Kaplan, J. O., Levis, S., Lucht, W., Sykes, M. T., Thonicke, K., and Venevsky, S.: Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ dynamic global vegetation model, Global Change Biology, 9, 161–185, https://doi.org/10.1046/j.1365-2486.2003.00569.x, 2003.

Snyder, J. P.: An Equal-Area Map Projection For Polyhedral Globes, Cartographica: The International Journal for Geographic Information and Geovisualization, 29, 10–21, https://doi.org/10.3138/27H7-8K88-4882-1752, 1992.

Strobl, C., Malley, J., and Tutz, G.: An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests, Psychological methods, 14, 323–48, https://doi.org/10.1037/a0016973, 2009.

Uber: Uber/H3: Hexagonal hierarchical geospatial indexing system, https://github.com/uber/h3, 2022.

Uher, V., Gajdo, P., Snasel, V., Lai, Y.-C., and Radecky, M.: Hierarchical Hexagonal Clustering and Indexing, Symmetry, 11, 731, https://doi.org/10.3390/sym11060731, 2019.

Wang, L., Ai, T., Shen, Y., and Li, J.: The isotropic organization of DEM structure and extraction of valley lines using hexagonal grid, Transactions in GIS, 24, 483–507, https://doi.org/https://doi.org/10.1111/tgis.12611, 2020.

Ware, C., Mayer, L., Johnson, P., Jakobsson, M., and Ferrini, V.: A global geographic grid system for visualizing bathymetry, Geoscientific Instrumentation, Methods and Data Systems, 9, 375–384, https://doi.org/10.5194/gi-9-375-2020, 2020.

Wolpert, D. H. and Macready, W. G.: An Efficient Method to Estimate Bagging's Generalization Error, Working papers, Santa Fe Institute, https://EconPapers.repec.org/RePEc:wop:safiwp:96-06-038, 1996.

World Bank, D. T.: Changes in Country Classifications, https://blogs.worldbank.org/opendata/changes-country-classifications, 2011.

Wright, J. W.: Regular hierarchical surface models: A conceptual model of scale variation in a GIS and its application to hydrological geomorphometry: A thesis submitted for the degree of doctor of philosophy at the University of Otago, Dunedin, New Zealand, Ph.D. thesis, University of Otago, 2019.

Wright, M. N. and Ziegler, A.: ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R, Journal of Statistical Software, 77, 1–17, https://doi.org/10.18637/jss.v077.i01, 2017.

Zohaib, M. and Choi, M.: Satellite-based global-scale irrigation water use and its contemporary trends, The Science of the total environment, 714, 136 719, https://doi.org/10.1016/j.scitotenv.2020.136719, 2020.