SPATIAL INCONSISTENCIES INFLUENCE GLOBAL ESTIMATES OF SOIL RESPIRATION:

HOW CAN WE IMPROVE GLOBAL PREDICTIONS?

by

Emma Stell

A thesis submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Master of Science in Geography

Fall 2020

© 2020 Emma Stell
All Rights Reserved

SPATIAL INCONSISTENCIES INFLUENCE GLOBAL ESTIMATES OF SOIL RESPIRATION:

HOW CAN WE IMPROVE GLOBAL PREDICTIONS?

by

Emma Stell

Approved: _____

Rodrigo Vargas, Ph.D.
Professor in charge of thesis on behalf of the Advisory Committee

Approved: _____

Tracy DeLiberty, Ph.D.
Chair of the Department of Geography and Spatial Sciences

Approved: _____

Estella Atekwana, Ph.D.
Dean of the College of Earth, Ocean, and Environment

Approved:

_____

Louis F. Rossi, Ph.D.
Vice Provost for Graduate and Professional Education and
Dean of the Graduate College

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Rodrigo Vargas. I would also like to thank my committee members, Dr. Ben Bond-Lamberty, Dr. Pinki Mondal, and Dr. Jing Gao. I am also grateful to members of the Vargas lab for their advice, help with code problems, and for being such a great group in general.

TABLE OF CONTENTS

LIST OF TABLES v
LIST OF FIGURES vi
ABSTRACT vii

Chapter

1 INTRODUCTION TO SOIL RESPIRATION 1

2 TOWARD IMPROVING ESTIMATES OF GLOBAL SOIL RESPIRATION 4

Introduction 4

Methods 9

Results 18

Discussion 28

3 CONCLUSIONS AND MOVING FORWARD 37

REFERENCES 40

Appendix

A SUPPLEMENTARY MATERIAL 48

LIST OF TABLES

Table 1 Rs zonal statistics……..24

Table 2 Rh zonal statistics. We used the Rh derived from the Bond-Lamberty (2004) equation to compare v3 to v5……27

LIST OF FIGURES

Figure 1 Measurement sites of both the SRDB v3 annual Rs observations, and the new measurement locations included within v5 that were utilized within this study……...11

Figure 2 Comparison of the model-derived mean and uncertainty between v3 and v5 of the SRDB. 2a and 2b are the mean and C.V. of v3, while 2c and 2d are the mean and C.V. of v5…….19

Figure 3 Difference maps comparing both the a) mean and b) C.V. of v3 vs v5. v3 values were subtracted from v5 values to get the resulting maps. Values less than zero indicate that v5 is underestimating in an area compared to v3, while values greater than zero indicate that v5 is overestimating in comparison to v3……..20

Figure 4 Global mean percent change (a) derived from the mean Rs difference map, along with the mean percent change plotted within climate space, i.e., the mean annual temperature and mean annual precipitation (b)…….22

Figure 5 Partial dependence plots of the six covariates used to model Rs. The x-axis represents the values of the respective covariates, while the y-axis represents the response of each variable within the QRF model…….23

Figure 6 Visualization of 3000 points chosen by the cLHS analysis to represent an idealized network…….25

Figure7 7a and 7b are from the model utilizing the sampling scheme from the cLHS analysis. Because there is no ground-truth data at those locations, we extracted the original model values from v5 to the cLHS locations. The resulting maps of the a) mean Rs and b) Rs C.V. are not based on ground-truth, but are merely a way to compare sampling schemes. 7c and 7d are from the model utilizing the sampling scheme from v5 of the SRDB. Although we do have ground-truth data at those locations, in order to compare the sampling scheme from Figure 7, we extracted the original model values from v5 to the SRDB point locations. The resulting maps of the c) mean Rs and d) Rs C.V. are not based on ground-truth, but are merely a way to compare sampling schemes…..26

ABSTRACT

Soil respiration (Rs), the efflux of CO_2 from soils to the atmosphere, is a major component of the terrestrial carbon cycle, but is poorly constrained from regional-to-global scales. The global soil respiration database (SRDB) is a compilation of in-situ Rs observations from around the globe that has been consistently updated with new measurements over the past decade. It is unclear whether the addition of data to new versions of the SRDB has produced better-constrained global Rs estimates. We compared two versions of the SRDB (v3.0 n=5173 and v5.0 n=10366) to determine how additional data influenced the global sum and uncertainty of Rs as predicted by a quantile regression forest model. A model parameterized using SRDBv3 yielded a global Rs sum of 88.6 Pg C yr^{-1}, and associated uncertainty of 29.9 (mean absolute error) and 57.9 (standard deviation) Pg C yr^{-1}, whereas parameterization using SRDBv5 yielded 96.5 Pg C yr^{-1} and associated uncertainty of 30.2 (mean average error) and 73.4 (standard deviation) Pg C yr^{-1}. SRDBv5 included new data from underrepresented regions (e.g., Asia, Africa, South America) which resulted in overall higher model uncertainty. Empirically estimated global heterotrophic respiration (Rh) from v3 and v5 were 49.9-50.2 (mean 50.1) and 53.3-53.5 (mean 53.4) Pg C yr^{-1}, respectively. The SRDBv5 is still biased towards northern latitudes and temperate zones, so we tested that an optimized global distribution of Rs measurements resulted in a global sum of 96.4 ± 21.4 Pg C yr^{-1} and lower model uncertainty. These results support current global estimates of Rs but highlight spatial inconsistencies that

influence model parameterization and interpretation that should be considered to improve predictions across the world.

Chapter 1

INTRODUCTION TO SOIL RESPIRATION

Soil respiration (Rs) is commonly known as the efflux of carbon dioxide (CO_2) from the soils to the atmosphere. It composes the second largest flux in the carbon cycle, after gross primary productivity (GPP) (Raich & Schlesinger 1992), the two of which are the major processes involved in land-atmosphere CO_2 exchange (Beer et al. 2010). 10% of the global atmospheric CO_2 cycles through the soils each year (Raich & Potter 1995), which is a magnitude ten times greater than anthropogenic emissions (Raich & Tufekcioglu 2000; Bond-Lamberty & Thomson 2010). Thus, Rs plays an important role in determining the carbon balance of terrestrial ecosystems. However, global estimates of Rs and other components of the terrestrial carbon cycle are poorly constrained and constitute a large portion of the uncertainty in earth system models such as CMIP4 and CMIP5 (Friedlingstein et al. 2006, 2014). Additionally, there are still large uncertainties in potential feedbacks in the face of rising atmospheric CO_2 (Friedlingstein et al. 2014).

Components of Rs include heterotrophic respiration (Rh), which is respiration through decomposition by microbes, and autotrophic respiration (Ra), which is maintenance respiration performed by the roots, rhizosphere, and mycorrhizae (Schlesinger & Andrews 2000; Bond-Lamberty & Thomson 2004; Ryan & Law 2005). Studies have attempted to establish a global relationship between Rs, Rh, and

Ra (Bond-Lamberty & Thomson 2004; Chen et al. 2014; Bond-Lamberty et al. 2018), but there are much fewer measurements of Rh than of Rs, making a global relationship difficult to establish.

Measurements of Rs began in the early 1900's (Smith & Brown 1931), with observations increasing throughout the last few decades, especially within the last few years (Bond-Lamberty & Thomson 2010, Jian et al. 2020a). Observations have historically been made manually, which is temporally sporadic, but recently there has been increased use of automated sampling that can provide continuous measurements at a single location (Vargas et al. 2011). Therefore, different measurement techniques for Rs have trade-offs between spatial and temporal resolution (Savage & Davidson 2003). To compensate for the temporal expectations of observing and calculating annual Rs for a site, it may be possible to use proxies such as the Rs at the mean annual temperature (MAT) of a site, which integrates the Q10 response to temperature (Bahn et al. 2010). Q10 refers to temperature sensitivity, and is a factor by which Rs is multiplied with every 10 degree increase in soil temperature (Davidson et al. 2006).

Complex drivers make it difficult to measure Rs across regional to global scales. The majority of studies investigating Rs at a global scale find high correlations between Rs measurements and mean annual temperature and precipitation (Singh & Gupta 1977; Raich & Schlesinger 1992; Raich & Potter 1995; Reichstein et al. 2003; Hibbard et al. 2005; Davidson et al. 2006; Reichstein & Beer 2008; Chen et al. 2010; Chen et al. 2014; Zhao et al. 2017). These are easily measured variables available at a global scale that can act as proxies for less easily measured drivers of Rs, such as soil

moisture availability. There are general high correlations with variables associated with site productivity (Raich & Schlesinger 1992; Reichstein et al. 2003; Hibbard et al. 2005; Chen et al. 2014) because primary production provides the organic fuel for metabolic activity (Raich & Schlesinger 1992). Vegetation affects Rs through the soil microclimate, detritus quality and quantity, and effects on root respiration. Therefore, shifting vegetation cover through land-use changes could have an impact on Rs rates (Raich & Tufekcioglu 2000; Hibbard et al. 2005).

There exist fewer global studies about Rs than any other terrestrial carbon flux (i.e., GPP, NPP) or greenhouse gas (i.e., methane, nitrous oxide) (Hashimoto et al. 2015). Recent global estimates have consistently higher values of annual global Rs flux than do earlier estimates (Schlesinger 1977; Raich & Schlesinger 1992; Raich & Potter 1995; Raich et al. 2002; Bond-Lamberty & Thomson 2010; Hashimoto et al. 2012; Chen et al. 2013; Hashimoto 2015; Xu & Shang 2016; Zhao et al. 2017; Adachi et al. 2017; Jian et al. 2018a).

The overarching goal of this study was to determine if adding new Rs measurements around the world improve global estimates of Rs. The objectives are: 1) model and map mean annual global Rs and its associated uncertainty at a 1 km resolution; 2) determine if the addition of recent Rs sites to a global database decreases the uncertainty of model predictions at the global scale; 3) investigate how model performance varies based on the distribution of sampling locations.

Chapter 2

TOWARD IMPROVING ESTIMATES OF GLOBAL SOIL RESPIRATION[1]

Introduction

Soil respiration (Rs) is the total efflux of CO_2 from soils to the atmosphere as a result of below-ground processes. Rs comprises the second largest flux in the carbon cycle, after gross primary productivity (GPP) (Raich & Schlesinger 1992), and is a major process involved in land-atmosphere CO_2 exchange (Beer et al. 2010). Roughly, 10% of the global atmospheric CO_2 cycles through the soils each year (Raich & Potter 1995), an order of magnitude greater than anthropogenic emissions (Raich & Tufekcioglu 2000). Two respiration processes make up Rs: autotrophic respiration (Ra) is the efflux of CO_2 due to plant roots, while heterotrophic respiration (Rh) is the efflux of CO_2 due to microbial decomposition of soil organic matter. Rh is an important component of Rs because it is strongly influenced by changes in soil temperature and moisture status, and long term changes could lead to increases of CO_2 in the atmosphere (Moinet et al. 2016). A rising Rh:Rs ratio could mean more soil organic carbon is being lost, but the sensitivity of Rh to a changing climate is largely unknown (Bond-Lamberty et al. 2018). It is therefore important to also investigate

[1]This chapter is under review (Stell, E. Warner, D., Jian, J., Bond-Lamberty, B., Vargas, R. (under review) Spatial inconsistencies influence global estimates of soil respiration: how can we improve global predictions?)

how Rh varies in comparison to Rs for a better overall understanding of global Rs processes.

Scientists have estimated Rs at the global scale for decades (Schlesinger 1977) and recent estimates converge to around 91-94 Pg C yr^{-1} (Bond-Lamberty 2018). These estimates are higher than global sums from earlier research that range from 68-76.5 Pg C yr^{-1} (i.e., Schlesinger 1977, Raich & Schlesinger 1992, Raich & Potter 1995, Raich et al. 2002). There is large variability in new estimates of Rs fluxes using different approaches, ranging from 66.6 - 108.6 Pg C yr^{-1} (Jian et al. 2018, Hursh et al. 2017), which introduces substantial uncertainty into climate forecasts. Over 40 Pg C yr^{-1} separate the highest and lowest global Rs estimates. The lowest estimate of 66.6 Pg C yr^{-1} was generated through an empirical model (Jian et al. 2018), while the highest estimate, of 108.6 Pg C yr^{-1}, was generated through a mechanistic model (Hursh et al. 2017). Historically, most global Rs estimates have been generated through empirical models (e.g., Schlesinger 1977, Raich & Schlesinger 1992, Raich & Potter 1995, Raich et al. 2002), with only some recent examples of mechanistic (e.g., Hursh et al. 2017) and machine learning approaches (e.g., Zhao et al. 2017, Jian et al. 2018, Warner et al. 2019).

It is technically impossible to directly measure large-scale Rs. Statistical models and machine learning (ML) offer a potential solution to this issue (Thessen 2016). Many recent studies have used ML (e.g., Zhao et al. 2017, Jian et al. 2018, Warner et al. 2019) which can help facilitate and improve large-scale estimates (i.e., global Rs) by relating Rs observations to ancillary vegetation, terrain, climate, and

survey data with global coverages. Some ML algorithms are particularly effective in modeling biogeophysical factors due to their ability to account for nonlinear relationships and low sensitivity to autocorrelation amongst predictor variables relative to more traditional linear models (Reichstein et al. 2019). ML approaches have been shown to outperform mechanistic and semi-empirical models in modeling carbon processes in which point-based observations are used for regional to global estimates (Reichstein et al. 2019), which shows promise for the use of ML algorithms in future environmental modeling and benchmarking (Bond-Lamberty 2018). Although ML offers a powerful tool for environmental prediction, it can be challenging to evaluate results, especially within ecological contexts (Basu et al. 2018).

Communicating uncertainty is an important aspect of properly reporting global Rs estimates. Uncertainties in land carbon cycle processes account for most of the uncertainty in overall CO_2 projections (Friedlingstein et al. 2006, 2014), but these models could be better constrained with more accurate estimates of global Rs. Many studies estimating global Rs report uncertainty, but not all. Out of early estimates, only one reports uncertainty, a standard deviation of ±4 Pg C yr^{-1} (Raich & Schlesinger 1992). The practice of reporting uncertainties became more commonplace starting in the early 2000s (Raich et al. 2002), and most studies of global Rs estimates since then have reported uncertainty (Bond-Lamberty & Thomson 2010, Chen et al. 2013, Hashimoto et al. 2012, 2015, Xu & Shang 2016, Zhao et al. 2017, Hursh et al. 2017, Warner et al. 2019). Current estimates of uncertainty vary widely, and one reason is

that uncertainty is reported in different ways, such as root mean square error (RMSE) (Hursh et al. 2017, Zhao et al. 2017), standard deviation (SD) (Raich & Schlesinger 1992), standard error (Bond-Lamberty & Thomson 2010, Chen et al. 2014), 95% confidence level (Xu & Shang 2016, Hashimoto et al. 2012, 2015), or a mixture of different measures (Warner et al. 2019, Chen et al. 2013, Raich et al. 2002). These may represent different uncertainty sources (e.g. year-to-year variability, measurement error, parametric uncertainty), so interpretation and comparison among these estimates should be done carefully (Enting et al. 2012).

Regardless of methodology, most global Rs modeling efforts rely on in situ measurements of Rs for calibration and/or validation (see however Konings et al. 2019). There have been several efforts to integrate point-based Rs measurements from studies around the world into publicly-accessible databases (Bond-Lamberty & Thomson 2010; Jian et al. 2018a; https://github.com/bpbond/cosore). The Soil Respiration Database (SRDB) is a database of annual Rs measurements from around the globe (Bond-Lamberty & Thomson 2010). The SRDB includes more than 10,000 total observations from over 2000 studies ranging from 1961-2017, making it a valuable resource for regional to global analysis. Version 4.0 of the SRDB was released in December 2018 with over 6500 observations from over 1400 studies, and Version 5.0 was released in April 2020 (Jian et al. 2020a) with a substantial increase in observations and spatial coverage. In addition, the Hourly and Daily Timescale Global Soil Respiration Database (HGRsD) is a database of hourly and daily Rs measurements also compiled from studies around the world (Jian et al. 2018a).

Finally, COSORE is a community database of continuous Rs measurements first released in June 2019, in which automated Rs measurements from individual sites are compiled into one database (Bond-Lamberty et al. 2020). With each subsequent version released, the spatio-temporal coverage of global Rs sites increases, facilitating additional insight to global Rs processes.

However, even with the expanded SRDB v5.0, there still exists an uneven point distribution of Rs observations around the globe. There are substantially more estimates of Rs in the Northern Hemisphere, especially in Europe, China, and North America (Xu & Shang 2016). Measurements are especially lacking in Africa (Epule et al. 2015), central Asia, and South America (especially in the Amazon). Most data points are located in temperate regions, with few data points in tropical and dryland ecosystems. This may lead to high uncertainty in spatial upscaling, given the role of tropical and dryland regions in regulating much of the global carbon cycle (Malhi & Grace 2000; Poulter et al. 2014). The uneven distribution of data points could bias regional-to-global representation and consequently the benchmarking of earth system models (ESMs). Improving ESMs requires extensive comparison between modeled and observed data, and updated observational data of Rs has been identified as a key area for improving (Hoffman et al. 2017). If the addition of more measurements decreases uncertainty in global estimates, then benchmarks in ESMs could be better constrained, enhancing the ability of researchers to accurately quantify global carbon fluxes. A critical question, then, is how much ‘value’ additional Rs observations (i.e., recently added to SRDBv5) are bringing to global Rs estimation in terms of decreasing

uncertainty and spatial bias. While previous research has argued for more observations to constrain models, there has been little investigation into how new Rs observations, and new observation locations, influence global Rs estimates.

The overarching goal of the study was to predict global estimates of Rs using two different versions of the SRDB (version 3.0 (v3) vs version 5.0 (v5)) in order to determine if adding new Rs measurements around the world improve global estimates of Rs. Our objectives are: 1) model and map mean annual global Rs and its associated uncertainty at a 1 km resolution; 2) determine if the addition of recent Rs sites to the SRDB decreases the uncertainty of model predictions at the global scale; 3) investigate how model performance varies based on the distribution of sampling locations. We hypothesize that SRDBv5 will have lower uncertainty in the overall global prediction of Rs as it has more sites and better distribution of information available for model parameterization. As the SRDB grows and expands, this is the first study comparing different versions of the database to investigate the effects of inclusion of additional measurements, especially with new additions across underrepresented regions of the world (e.g., Africa, South America, Asia).

Methods

Rs training data

We used data from two versions of the global soil respiration database (SRDB): v3 (downloaded September 5, 2018; http://dx.doi.org/10.3334/ORNLDAAC/1235) and v5 (downloaded April 9, 2020;

https://github.com/bpbond/srdb). The SRDB is an initiative to compile field studies of Rs from around the globe into one harmonized database (Bond-Lamberty and Thomson 2010). Information selected from the database includes record number, study number, latitude, longitude, manipulation, measurement method, annual Rs, biome, and study midyear. Data were only included if certain criteria were met: Rs_annual < 10,000 g C m^{-2} yr^{-1} (accounting for outliers); there were no field manipulations; and the measurement methods were either IRGA or gas chromatography.

A total of 2321 data points from 618 studies for v3, and 4115 data points from 1036 studies for v5 of the SRDB, met these requirements. Consequently, our analyses using v5 included 44% more training data points. In comparison to v3, the expanded v5 database has increased spatio-temporal coverage, newly included studies within Chinese and Russian scientific literature, and enhanced interoperability with other carbon datasets (Jian et al., 2020)

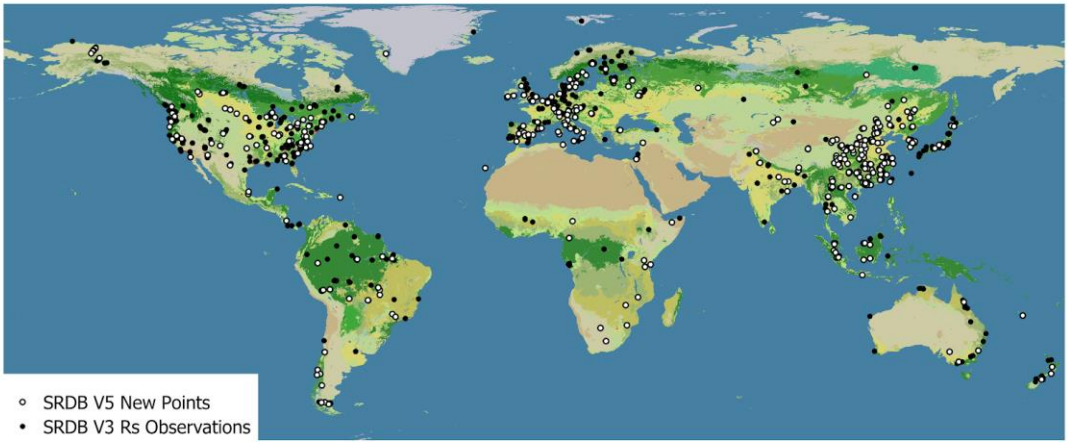


Figure 1. Measurement sites of both the SRDB v3 annual Rs observations, and the new measurement locations included within v5 that were utilized within this study.

Prediction factors for global Rs

Our initial predictors are a set of 15 globally-distributed predictors that were represented by climatological, vegetative, and geophysical predictive factors related to Rs. Prediction factors are represented by environmental information from worldgrids.org (last accessed Jan. 2018), an initiative of ISRIC Soil Information (Reuter & Hengl 2016, and from worldclim.org (Fick & Hijmans 2017). The *worldgrids* and *worldclim* predictors come from three sources: remote sensing, climate surfaces, and digital terrain analysis (see complete list and description of layers in Supplementary Material). These were mean annual temperature, mean annual

precipitation, mean of the monthly MODIS Enhanced Vegetation Index time series data, mosaic vegetation (grassland/shrubland/forest) (50-70%) / cropland (20-50%), artificial surfaces and associated areas (Urban areas >50%), geological ages based on the surface geology, land cover classes based on the Medium Resolution Imaging Spectrometer (MERIS) full resolution images, mean potential incoming solar radiation derived in the System for Automated Geoscientific Analyses (SAGA) GIS, mean value of the 8-day MODIS LAI time series data, Long-term precipitation for periods Nov/Dec/Jan, Feb/Mar/Apr, May/Jun/Jul, and Aug/Sep/Oct based on WorldClim.org data, Slope map in percent derived using the DEMSRE3, and SAGA Topographic Wetness Index derived using the DEMSRE3. The predictors are all the same spatial resolution and projection, and were subsequently organized into a regression matrix for further analysis. The regression matrix was used for variable selection and subsequent modeling as described in section 2.3.

Prediction of global Rs and model evaluation

We use VSURF in order to refine and select the most relevant environmental variables for prediction. The two main steps composing VSURF are interpretation and prediction. First, predictor variables are evaluated for their individual impact on model performance, then predictors are analyzed for redundancy to find a smaller set of variables sufficient for prediction (Genuer et al. 2010, 2015). In the interest of comparing the same model across the two versions of the SRDB, we used the

variables selected for v5 of the SRDB by VSURF for predicting Rs using SRDB v3 and v5.

Following Warner et al. 2019, we use quantile regression forests (QRF) to predict Rs at 1 km resolution at the global scale (see Objective 1). QRF is a generalization of Random Forests (RF), which is a commonly used machine learning algorithm (Biau & Scornet 2016; Breiman 2001). Unlike RF, in QRF the conditional prediction distributions are preserved, allowing users to infer quantiles and other distribution statistics beyond the conditional mean (Meinshausen 2006). Furthermore, QRF has been shown to outperform Regression-Kriging in predictions of uncertainty and can be used when data is sparse (Vaysse & Lagacherie 2017). The QRF model was trained using the environmental covariates identified by VSURF, and using 2321 training SR data for v3 and 4115 observations for v5. Model parameters *ntree* (the number of trees grown in the QRF model) and *mtry* (the number of variables considered at each split in the prediction trees) were calibrated using five repetitions of a 10-fold cross validation. The resulting model was then extrapolated to the set of global environmental covariates at 1 km resolution in order to generate conditional prediction distributions of Rs for each pixel. The mean of the conditional distribution of the QRF was used to make the final global mean Rs prediction map using SRDB v3 or v5. We used the conditional SD (standard deviation) and C.V. (coefficient of variation) for looking at model prediction consistency within each pixel.

Global Rs estimates were made as the sum of the mean of the conditional prediction distribution at each pixel multiplied by its surface area and compared

among v3 and v5. Uncertainty was reported as the sum of the SD of the conditional prediction distribution at each pixel multiplied by its surface area (ε_{SD}), the sum of the model RMSE multiplied by the area of each pixel (ε_{RMSE}), and the sum of the model MAE (mean absolute error) multiplied by the area of each pixel (ε_{MAE}) (see Objective 2). Model prediction consistency at each pixel was expressed as C.V., which is the standard deviation of the conditional distribution of the QRF model divided by the mean. To illustrate the discrepancies between the model outputs, we created difference maps by subtracting Rs mean or C.V. of v3 from v5. Zonal statistics were extracted from our annual Rs prediction raster for 16 IGBP vegetation classes based on MODIS data (Channan et al. 2014) in order to compare the mean and total Rs sums among them.

We examined the effects of each predictor variable on model output through partial dependence plots to better understand the inner workings of our QRF model (Figure 5). Partial dependence plots show the general response of model output (i.e., predicted Rs) over the range of an individual independent variable over the marginal distribution of all other independent variables (Goldstein et al. 2015). Partial dependence plots are useful for interpreting relationships between each individual predictor and the model output, while accounting for the effects of the other predictors within the model (Greenwell 2017). This partially accounts for the drawback that machine learning methods can be more difficult to interpret relative to more traditional methods. Within the plots, positive or negative trends indicate a general positive or negative influence on predicted output among training instances, while a flat trend

indicates that the effects of the predictor on the model output are weak, or inconsistent across individual training instances.

Alternative global measurement locations

We applied a framework to identify potential sites that better represent the covariate space of Rs across the world. This analysis aims to inform potential biases in the SRDB and identify measurement locations around the world to better represent global Rs spatial variability (see Objective 3). Individual measurements within the SRDB were never planned to be representative of global Rs patterns, but because of the wealth of information compiled, the SRDB is applicable to address research at the global scale (Bond-Lamberty et al. 2018; Bond-Lamberty and Thomson 2010). That said, the database is and has been biased towards studies in northern latitudes and temperate zones. Even with an additional 1794 site measurements included in China, Russia, and other underrepresented regions, SRDB v5 still has limitations in global spatial and ecological representation (Jian et al. 2020a).

We used a conditioned Latin Hypercube Sampling (cLHS) analysis to identify alternative potential sites across the globe, modeled after similar analyses (e.g. improving the design and representation of environmental networks (Villarreal et al. 2019)). cHLS is a sampling strategy in which *x* new sites are chosen such that the multivariate distribution is maximally stratified (Minasny & McBratney 2006). While cLHS has mostly been used within digital soil mapping (e.g., Chu et al. 2010, Mulder et al. 2013, Silva et al. 2015, Stumpf et al. 2016), some studies have utilized cLHS for

other environmental applications (e.g., Lin et al. 2009, Yin et al. 2016, Yin et al. 2017, Contreras et al. 2019, Villarreal et al. 2019). We applied the cLHS algorithm to choose 3000 sites around the globe (because SRDBv5 has >3000 data entries) in locations that maximize the representation of the multivariate space of the previously selected environmental covariates used to model Rs.

To test whether an optimized sampling design would decrease uncertainty for global Rs, we ran two Rs models with different distributions of data points. The ability of statistical models to well predict a given variable relies not only on the quantity of available data, but also on how well that data represents the variability of the space in which predictions are being made. For a comparable test of sampling schemes, we used our continuous v5 map as the source of input for Rs values. First, we took the point locations identified by the cLHS analysis, and extracted the values from our v5 global map at the locations of these 3000 points. Next, we extracted the values from the locations of the SRDB points for v5 from the same v5 global map. The point locations from these different sampling schemes were run in our model for comparison of the mean spatial distribution of the QRF model and prediction CVs.

The resulting maps are useful to test how the spatial distribution of the training data (i.e., data available in the SRDB) influence Rs estimates and uncertainty around the world. However, these maps are not intended to be accurate predictions of global Rs, as their inputs are the previous model outputs of our global Rs estimates using v3 and v5. Information about the influence of spatial distribution on global Rs estimates is important for interpretation of the model outputs using the original SRDB v3 and

v5, and provides insights about where new measurements are needed and model mismatches.

Calculation of Rh

We calculated heterotrophic respiration (Rh) on a pixel-by-pixel basis from our rasters of global estimated Rs. Rh represents the portion of Rs contributed by microbes and other heterotrophic organisms within the soil (Chapin et al. 2006). The role of Rh is directly relevant to issues of carbon balance, soil carbon stability, and climate feedbacks (Bond-Lamberty et al. 2018). Therefore, Rh was calculated from global Rs (at 1x1 km pixels) from SRDB v3 and v5 for comparison. The following equations were used to estimate Rh:

$$\ln(Rh) = 1.22 + 0.73\ln(Rs)$$

(Bond-Lamberty et al. 2004)

$$Rh/Rs = -0.138\ln(Rs) + 1.482$$

(Subke et al. 2006)

These simple functions provide a useful and generally robust (for undisturbed systems) approximation of Rh, although we note that the increasing availability of direct Rh measurements means that global-scale upscaling for this flux is now possible (Tang et al. 2020).

Code and data availability

All analyses were completed using R (R Core Team 2019). Packages raster, VSURF, quantregForest, pdp, and cLHS were used. Data products will be available at the ORNL DAAC (data submission is being evaluated by ORNL DAAC).

Results

Prediction factors for global Rs

Six variables had the highest impact on the prediction of Rs. These variables were mean annual temperature (MAT), mean annual MODIS enhanced vegetation index (EVI), mean precipitation from November to January (WMP), mean precipitation from February to April (SpMP), mean precipitation from May-July (SMP), and mean precipitation from August-October (FMP). These six variables were used for the prediction of Rs with both the v3 and v5 SRDB.

Prediction of global Rs and model evaluation

We calculated the mean sum, as well as the ε_{SD}, ε_{RMSE}, and ε_{MAE} of both v3 and v5. The best *ntree* and *mtry* QRF parameters were 210 and 1, respectively. For v3 we found a final model R^2, RMSE, and MAE of 0.61, 368.8 g C m^{-2} yr^{-1}, and 200.8 g C m^{-2} yr^{-1}, respectively, while for v5, we found a final model R^2, RMSE, and MAE of 0.61, 357.0 g C m^{-2} yr^{-1}, and 202.4 g C m^{-2} yr^{-1},-respectively. The sum of the mean of the conditional distribution of the QRF model for global Rs with v3 data was 88.6 Pg C yr-

$^{-1}$, while the sum of the mean for v5 was 96.5 Pg C yr^{-1}(Figures 2a and 2c). For v3, the ε_{SD}, ε_{RMSE}, and ε_{MAE} was 57.9, 54.9, and 29.9 Pg C yr^{-1}, respectively. For v5, the ε_{SD}, ε_{RMSE}, and ε_{MAE} was 73.4, 53.2, and 30.2 Pg C yr^{-1}, respectively. The mean global C.V. was 72.2 +- 21.8% for v3, and 89.8 +- 40.4% for v5.

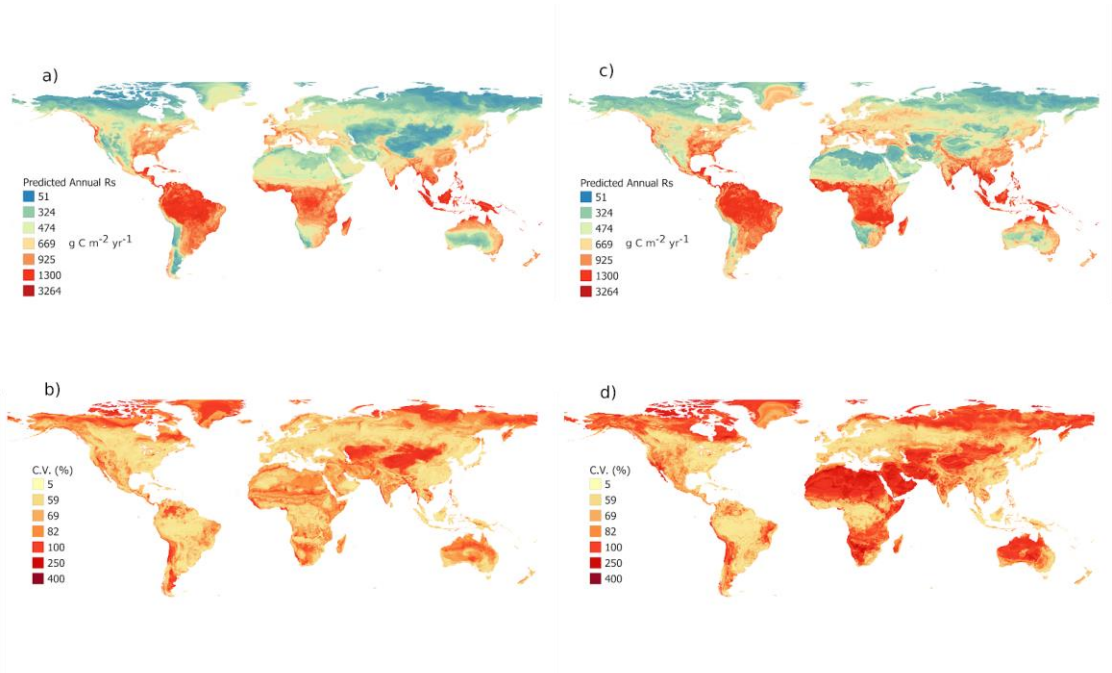


Figure 2. Comparison of the model-derived mean and uncertainty between v3 and v5 of the SRDB. 2a and 2b are the mean and C.V. of v3, while 2c and 2d are the mean and C.V. of v5.

Difference maps illustrate the discrepancy between the model outputs of v3 vs v5 (Figure 3). For the mean difference (Figure 3a), values are especially high in northern central Asia, eastern Europe, western semi-arid United States, portions of central Africa, and southern Australia.

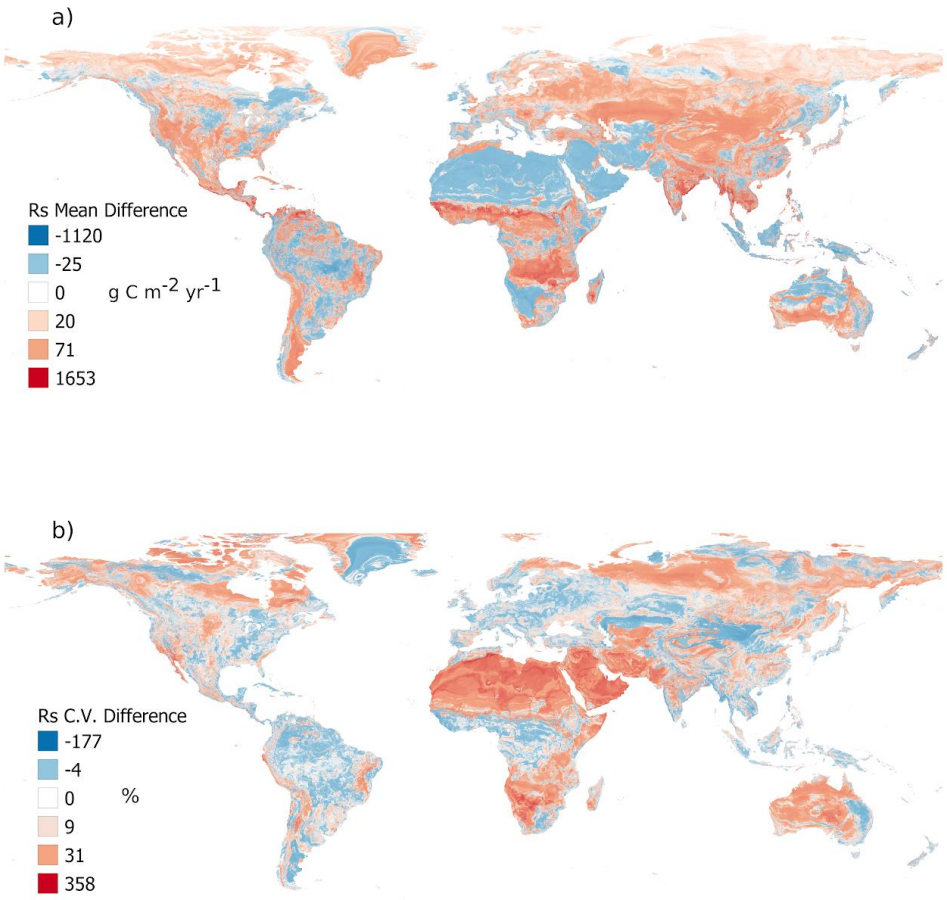
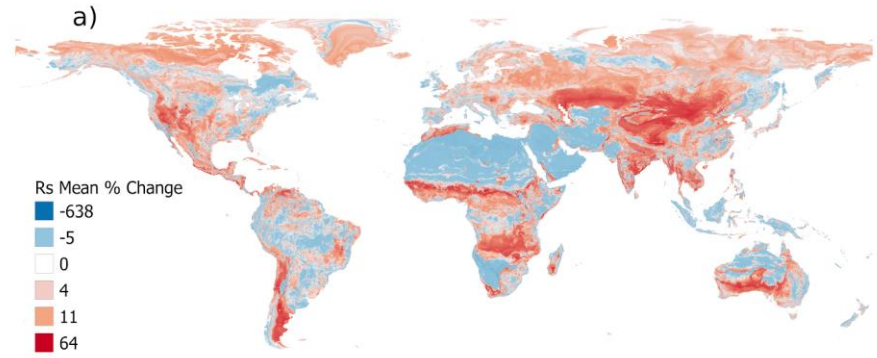


Figure 3. Difference maps comparing both the a) mean and b) C.V. of v3 vs v5. v3 values were subtracted from v5 values to get the resulting maps. Values less than zero indicate that v5 is underestimating in an area compared to v3, while values greater than zero indicate that v5 is overestimating in comparison to v3.

The mean percent changes between v3 and v5 were highest in arid/semi-arid regions (Figure 4a). Within different regions of the world, the same numerical change

can be a large percentage change within a region with typically low values, but a small percentage change within a region with higher values. The values of the mean percent change range from -638-66.4%, with a normal distribution (Figure 4b). A subsample of percent differences were plotted within climate space (i.e., along the global distributions of MAP and MAT (Figure 4b)) demonstrating higher values of Rs in SRDBv5 in wet areas (>200 mm of annual precipitation)
around the world.



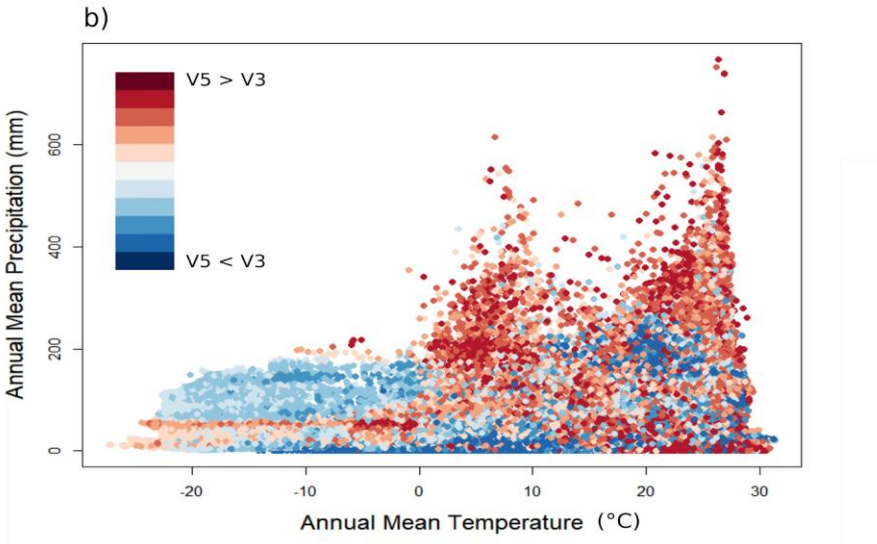


Figure 4. Global mean percent change (a) derived from the mean Rs difference map, along with the mean percent change plotted within climate space, i.e., the mean annual temperature and mean annual precipitation (b).

v3 and v5 produced similar partial dependence plots for each variable, with the exception of FMP (mean fall precipitation), which showed a stronger response to v5, and SpMP (mean spring precipitation), which showed a stronger positive effect to v3 within lower values. MAT shows a continual strong positive response to the model, indicating large dependency for the model over the range of its values. Only higher EVI values (>3000) exhibited a strong dependency on the model. The four precipitation variables show a stronger response in low-mid amounts of precipitation, with the response dropping or leveling out for values >300-400mm.

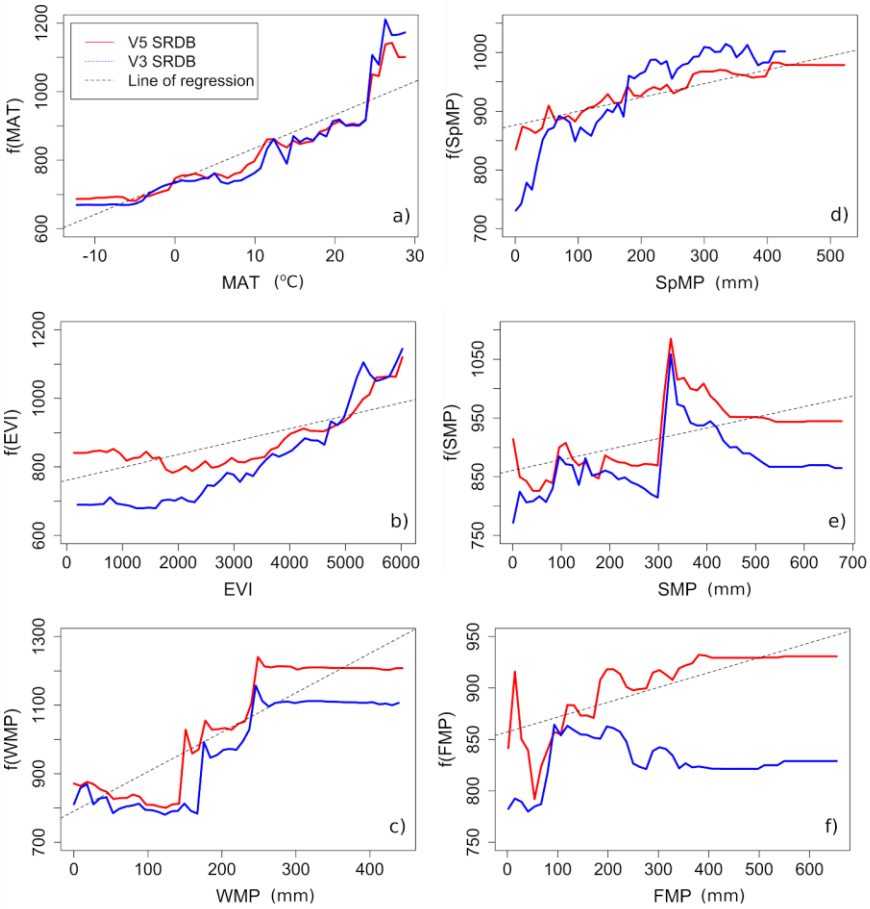


Figure 5. Partial dependence plots of the six covariates used to model Rs. The x-axis represents the values of the respective covariates, while the y-axis represents the response of each variable within the QRF model.

Model estimated Rs varied widely across the 16 different IGBP classes throughout the world, but Evergreen Broadleaf Forests (EBF) show the highest mean Rs. Table 1 shows the Rs mean, total sum, and C.V. for each IGBP class for v3 and v5 of the SRDB. Mean Rs in EBF was 1317.66 g C m^{-2} for v3, and 1305.22 g C m^{-2} for v5. Savannas had the next largest mean Rs, at 947.73, and 1076.73 g C m^{-2} for v3 and v5, respectively.

Table 1. Rs zonal statistics

LUName	LUShort	% land cover	v5 Mean (g C m-2)	v5 Sum (Pg C yr-1)	v3 Mean (g C m-2)	v3 Sum (Pg C yr-1)	v5 CV Mean (%)	v3 CV Mean (%)
Evergreen Needleleaf Forest	**ENF**	3.59	565.18	2.17	545.08	2.09	71.92	61.70
Evergreen Broadleaf Forest	**EBF**	7.72	1305.22	18.01	1317.66	18.19	55.78	58.22
Deciduous Needleleaf Forest	**DNF**	1.64	388.65	0.59	323.41	0.49	85.89	69.40
Deciduous Broadleaf Forest	**DBF**	0.88	884.54	1.17	862.70	1.14	55.55	52.97
Mixed Forest	**MF**	8.39	687.52	6.9	631.90	6.42	59.23	53.55
Closed Shrublands	**CSh**	0.05	707.98	0.06	706.21	0.06	86.14	71.19
Open Shrublands	**OSh**	19.01	401.98	9.38	352.00	8.41	103.30	82.06
Woody Savannas	**WSv**	7.47	858.75	10.25	724.27	8.53	79.83	68.23
Savannas	**Sv**	5.53	1076.73	10.44	947.73	9.2	84.77	69.43
Grasslands	**GL**	14.84	544.07	10.78	438.51	8.94	91.47	78.19
Permanent Wetlands	**PW**	0.89	587.54	0.71	554.03	0.66	90.20	67.09
Croplands	**Crop**	8.37	757.52	9.07	670.42	8.01	67.06	60.29
Urban and Built-Up	**Urb**	0.26	820.63	0.3	720.40	0.27	71.06	65.81
Cropland/Natural Mosaic	**Cr/Nat**	4.90	955.38	7.34	833.56	6.4	63.12	60.75
Permanent Snow and Ice	**Snow**	4.83	482.92	1.38	377.56	1.08	103.29	96.34
Barren or Sparsely Vegetated	**Bare**	11.64	368.60	6.88	401.47	7.66	149.94	83.71

Prediction of new measurement sites

The cLHS analysis evenly distributed 3000 points around the globe to represent an idealized network (Figure 6). The Rs values predicted by our QRF model

at these point locations were used in our original QRF model to create a map based on these idealized points (Figure 7). While the sum of this model output is similar, at 96.4 Pg C yr^{-1}, the total sum SD and C.V. are much lower, with a sum SD of 21.4 Pg C yr^{-1}, and a mean C.V. of 24.8 +- 10.1%.

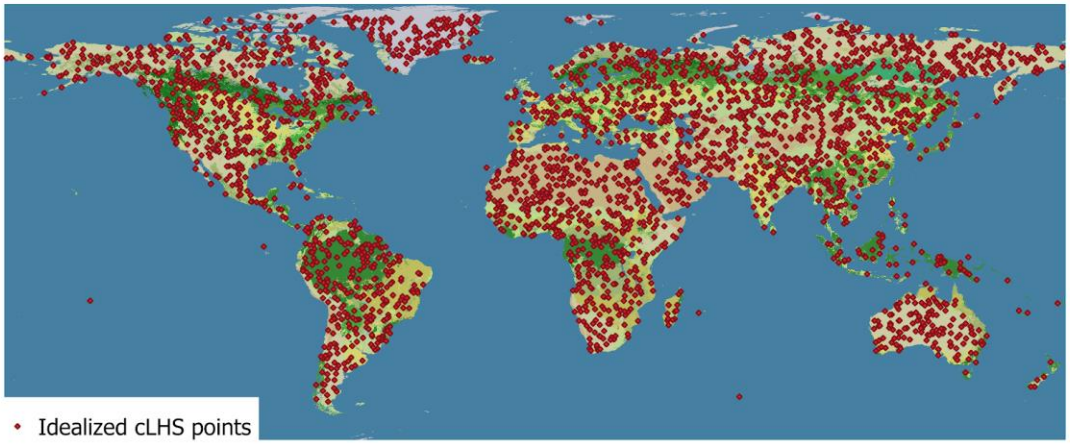


Figure 6. Visualization of 3000 points chosen by the cLHS analysis to represent an idealized network.

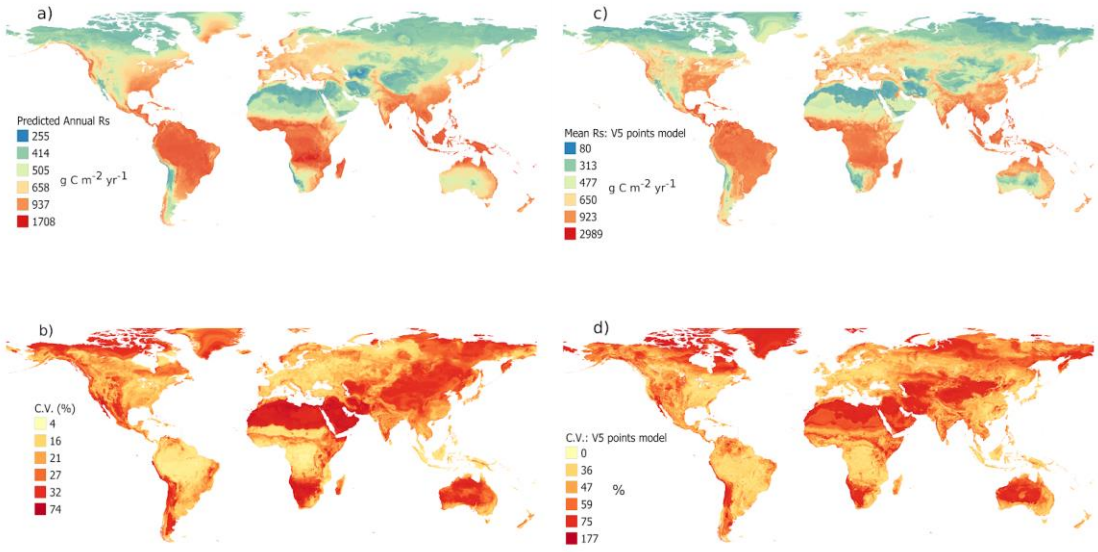


Figure 7. 7a and 7b are from the model utilizing the sampling scheme from the cLHS analysis. Because there is no ground-truth data at those locations, we extracted the original model values from v5 to the cLHS locations. The resulting maps of the a) mean Rs and b) Rs C.V. are not based on ground-truth, but are merely a way to compare sampling schemes. 7c and 7d are from the model utilizing the sampling scheme from v5 of the SRDB. Although we do have ground-truth data at those locations, in order to compare the sampling scheme from Figure 7, we extracted the original model values from v5 to the SRDB point locations. The resulting maps of the c) mean Rs and d) Rs C.V. are not based on ground-truth, but are merely a way to compare sampling schemes.

The distribution of points chosen by the cLHS analysis would decrease the C.V., and is not just an effect of the model. The global sum for the model utilizing v5 point locations was slightly higher than the model utilizing the cLHS points, at 96.1 Pg C yr^{-1}. The SD and mean C.V. were higher than in the model utilizing the cLHS points, with values of 45.4 Pg C yr^{-1}, and 55.4 +- 21.0%, respectively.

Calculation of Rh

For v3, Rh was calculated to be 49.9-50.2 Pg C yr^{-1} (mean 50.1 Pg C yr^{-1}), while for v5 Rh was calculated to be 53.3-53.5 Pg C yr^{-1}(mean 53.4 Pg C yr^{-1}). Figures can be found in Supplementary Material. There was an increase in Rh from v3 to v5 in northern temperate regions, central and eastern Asia, Australia, and dryland areas in North and South America. Many areas of the world that experienced increases in Rh fall in arid/semi-arid areas. We calculated the zonal statistics for v3 and v5 Rh based on the Bond-Lamberty (2004) equation (Table 3). The largest Rh sums for both v3 and v5 were found in Evergreen Broadleaf Forest, Grasslands, and Open Shrublands. The lowest Rh sums were in Closed Shrublands, Urban and built up areas, and Deciduous Needleleaf Forest.

Table 2. Rh zonal stats. We used the Rh derived from the Bond-Lamberty (2004) equation to compare v3 to v5.

LUName	LUShort	% land cover	v5 Mean (g C m-2)	v5 Sum (Pg C yr-1)	v3 Mean (g C m-2)	v3 Sum (Pg C yr-1)
Evergreen Needleleaf Forest	**ENF**	3.59	342.12	1.3	333.24	1.27
Evergreen Broadleaf Forest	**EBF**	7.72	633.26	8.74	637.58	8.80
Deciduous Needleleaf Forest	**DNF**	1.64	260.86	0.39	227.86	0.35
Deciduous Broadleaf Forest	**DBF**	0.88	477.13	0.63	469.32	0.62
Mixed Forest	**MF**	8.39	395.93	3.95	371.82	3.74
Closed Shrublands	**CSh**	0.05	404.24	0.03	404.06	0.03

Open Shrublands	**OSh**	19.03	264.73	5.99	239.07	5.50
Woody Savannas	**WSv**	7.47	455.46	5.27	405.51	4.65
Savannas	**Sv**	5.54	545.43	5.27	498.29	4.81
Grasslands	**GL**	14.80	330.47	6.42	279.10	5.53
Permanent Wetlands	**PW**	0.89	342.79	0.39	328.33	0.37
Croplands	**Crop**	8.38	424.24	5.04	388.88	4.62
Urban and Built-Up	**Urb**	0.25	448.78	0.17	411.35	0.15
Cropland/Natural Mosaic	**Cr/Nat**	4.90	498.77	3.79	453.09	3.44
Permanent Snow and Ice	**Snow**	4.82	302.30	0.85	253.32	0.71
Barren or Sparsely Vegetated	**Bare**	11.66	250.23	4.68	266.62	5.06

Discussion

Global Rs and Rh predictions

The global Rs and Rh sum increased from v3 to v5. The global Rs and Rh sums for v3 were 88.6 and 50.1 Pg C yr^{-1}, and for v5 were 96.5 and 53.4 Pg C yr^{-1}, respectively. There was an 8.9% increase in Rs and a 6.9% increase in Rh between v3 and v5 of the SRDB. These predictions hover just outside the range of other recent global estimates, which have been found to be converging on a global Rs of 91-94 Pg C yr^{-1} (Bond-Lamberty 2018). Our Rs predictions agree well with other RF implementations (Jian et al. 2018; Warner et al. 2019) based on the SRDB (Bond-Lamberty & Thomson 2010). Our Rh predictions compare well to other empirical global Rh estimates of 50 Pg C yr^{-1} (Raich & Schlesinger 1992) and 49.7 Pg C yr^{-1} (Warner et al. 2019), while falling within the range estimated by ESMs (i.e., 41.3-71.6 Pg C yr^{-1}) (Shao et al. 2013).

The six overall predictors selected (using an unsupervised approach) to predict Rs agree well with other studies. We highlight that it is critical to properly select environmental predictors within a machine learning framework to improve model interpretability, reduce model overparameterization, and reduce computing costs (i.e., learning time). Our unsupervised approach (VSURF) maximized model parameterization while reducing the covariate space to improve model interpretability. Overall, temperature and precipitation are well-known predictors of local-to-global Rs, and other studies have shown the importance/influence of productivity variables on Rs (e.g., Hursh et al. 2017). We investigated the effect of each predictor on modeled Rs values via partial dependence plots. MAT showed a weak response for lower temperatures, but an overall positive response throughout the rest of its range of values, only dropping off in the highest temperatures. MAT is known to be a suitable predictor of annual Rs across different ecosystems (Bahn et al. 2010). EVI showed little effect on Rs within lower values, but an increasingly positive trend emerged in the middle to upper ranges of its values. This could be because higher vegetation indices (such as EVI) is related to increased substrate supply, which directly influences Rs (Davidson et al. 2006). The effects of precipitation variables were more varied. SpMP showed the most consistent positive trend across the range of its values before its influence declined after ~400mm. FMP only showed influence between ~50-200mm, while WMP showed the most influence from 150-250mm. SMP only showed consistent influence between 300-350mm. The contrasting influences of the precipitation variables across different ranges of their values could explain why

VSURF selected seasonal precipitation variables instead of annual mean precipitation. The biggest differences in the pdp's between v3 and v5 are in SpMP and FMP. This is possibly due to the large increase in observations in regions with different precipitation seasonality, such as in regions of Asia that are influenced by monsoons. It is worth noting that just because the pdp's do not show a steep slope does not mean that there is no effect; there could be significant noise effects across the different trees that cancel each other out.

Contrary to our hypothesis, model uncertainty did not decrease when using SRDB v5 data over v3 data. Between the two versions of the SRDB, there was an increase of 1794 usable training data points (~77%). We expected that with an increased number of measurements, uncertainty should decrease in v5. The model uncertainty metrics (RMSE and MAE) were similar between v3 and v5. However, both SD and C.V. increased between v3 and v5. The SD sum increased by 15.5 Pg C yr^{-1} (a 26.8% increase), while the mean C.V. increased by 17.6%. This is not a reflection of the accuracy of the model, but is instead a result of the effect on consistency in model runs by adding more data points, especially across underrepresented regions around the world (i.e., areas with few data points). In other words, the model is biased with information from northern latitudes but has inconsistencies (i.e., larger uncertainty) when new but limited information is added in underrepresented areas. For v3, the C.V. is lower because the model has consistency across some areas even without data assuming results from a similar multivariate space. While the SD and C.V. saw substantial increases, the MAE and RMSE were

comparable between v3 and v5. The MAE and RMSE are generated through the 10-fold cross-validation, which evaluates the model in comparison to known data. Therefore, the model capacity to represent the SRDB is mostly similar between v3 and v5. However, the SD, and the C.V. from which it is derived, evaluate the consistency of the model runs. The effect of increasing data is to decrease the model consistency with greater variability among the regression forest predictions. Essentially, the training of the model is similar among versions, but the variability of the model prediction is higher in v5.

Many large differences in mean Rs were located in arid/semi-arid regions. The difference maps (Figure 3) and percent change map (Figure 4a) visually illustrate the differences between v3 and v5. Many arid/semi-arid areas, especially in the eastern hemisphere, experienced decreases in Rs from v3 to v5, including the Middle East, northern Australia, Saharan Africa, and southern Africa. Of these regions, only southern Africa had an increase in data points from v3 to v5. Even with the decreases observed in Figure 4a, only two IGBP land cover classifications saw decreases in mean and sum Rs, evergreen broadleaf forests and barren or sparsely vegetated land cover. In contrast to the eastern hemisphere, in the western hemisphere there was an increase in Rs in many dryland ecosystems, such as in the western United States, and South American arid/semi-arid regions. There was an increase in data points in both of these regions. Other regions with an increase in Rs include Central America, eastern Europe, Greenland, southern Australia, and in sub Saharan African regions dominated by dryland ecosystems (i.e., grasslands, savannas, and woody savannas) (Figure 3a). It

is possible that the large areas of contiguous woody savannas in sub saharan Africa could explain the large increase in mean and sum Rs (134.48 g C m^{-2} yr^{-1} and 1.72 Pg C yr^{-1}, respectively) between v3 and v5. While no data points were added in this region with v5, data points were added in other regions (i.e., eastern China) that are within woody savannas. Both the mean and sum Rs in savannas increased from v3 to v5 (by 128.99 g C m^{-2} yr^{-1} and 1.24 Pg C yr^{-1}, respectively), possibly because of the increase in the number of points within underrepresented savannas in Africa and South America.

The difference map for C.V. confirmed that there were large differences within dryland areas, including northern Africa, southern Africa, the Middle East, Australia, Pacific coastal areas in South America, and high latitudes. Many of these areas are located within barren or sparsely vegetated land cover, which corresponds to the 66.24% increase in C.V. Given the large increase in C.V. within this land cover type, measurements are especially needed in these areas. There was a decrease in C.V. values from v3 to v5 in most of Greenland, central Europe, the Amazon, and central Africa within tropical regions (Figure 3b). The only land cover classification with a decrease in C.V. was evergreen broadleaf forest, corresponding to the aforementioned tropical regions. Within IGBP classifications, there were higher differences in C.V. within permanent wetlands, barren areas, and open shrublands. Dryland environments such as the latter two are large drivers of interannual variability in Rs, with the majority of variability being attributed to the Southern Hemisphere (Poulter et al. 2014, Ahlstrom et al. 2015). Accurately accounting for large scale Rs fluxes in these ecosystems remains a challenge for global carbon cycle research.

The mean percent change (Figure 4a) had the largest increases in grasslands, open shrublands, savannas, and in Asian barren regions, reinforcing that drylands are large sources of variability. As expected, the percent change highlighted regions that were not elucidated within the mean difference map, mainly within arid/semi-arid regions. Within the global climate space (Figure 4b), v5 had consistently higher values than v3 in warmer, wetter (precipitation above ~200-250mm yr^{-1}) climates. On the other hand, v5 had lower values than v3 in colder, arid (precipitation below ~200-250mm yr^{-1}) climates. These differences on opposite ends of the climate spectrum indicate that new measurements should be focused in these climate spaces.

Where to measure for improving global representativeness

Even with improvements in spatial coverage with each subsequent version of the SRDB, it is still biased toward the northern hemisphere and temperate ecosystems (Jian et al. 2020a). The majority of the world's landmass is located in the northern hemisphere, along with more economically developed nations. The bias toward the northern hemisphere and more temperate regions affects global Rs estimates by influencing estimations of Rs across multivariate spaces with little data to parameterize models. In other words, the model will assume a value of Rs relatively similar to the closest multivariate space (likely form a northern hemisphere) when no data is available. We report that the effect of adding more data for model parameterization from v3 to v5 resulted in an increase in global Rs sum of nearly 8 Pg C yr^{-1}. This is not to say that increasing the number of measurement sites is not needed,

but our results highlight that we as a community should be strategically targeting specific areas for measurement and better constrain global estimates.

Not every point identified by the cLHS algorithm is able to realistically be measured. Many regions are poorly accessible due to lack of infrastructure, extreme climate, funds, remoteness, or a combination of these. However, the even distribution of points indicates that better spatially distributed measurement sites would better represent the global climate space. A previous study found that if Rs measurements were randomly distributed around the globe, only 591 measurements are needed to have a 5% relative error at a 95% confidence level- even less for a 10% relative error (Xu & Shang 2016). Many of the points identified from the cLHS analysis were located within the Sahara Desert and within the Australian outback, arid regions that show little heterogeneity. Although arid lands have lower Rs fluxes, they cover 30% of the world, representing a large multivariate space within the earth system. It is still unknown whether more points are needed in regions with less spatial variability, even if not adding points could lead to increased bias within the multivariate space. It has been recently found that Rs measured when soil temperature or air temperature reaches its MAT is able to well represent annual Rs (Jian et al. 2020b). This indicates that for those inaccessible regions a combination of meteorological data with Rs rates at MAT could decrease uncertainty in global Rs estimates, potentially overcoming some of the logistical challenges associated with data collection in data sparse areas.

Our models using the cLHS points vs the v5 points are a proof of concept that an optimal experimental design would decrease uncertainty in global Rs sums. The

cLHS model and the v5 points model had very similar global sums, of 96.4 and 96.1 Pg C yr-1, respectively. However, both the cLHS model SD sum and mean C.V. were less than half that of the v5 points model. While the v5 points model has a lower SD than the original modeled value, this is expected as the v5 points model input values were from the original model output. Looking at the maps of the cLHS model (Figure 7) and the v5 points model (Figure 8), the mean Rs and C.V. global patterns are similar between the two figures. This is also expected, as they are based on the same original model. However, the vast decrease in uncertainty in the cLHS model shows that better distributed input data would decrease the uncertainty in current models.

Moving forward, it is important to reduce uncertainty in global Rs models. Our proof of concept models showed that we can have greater confidence in the model outputs of Rs upscaling efforts by improving representation of underrepresented areas with a better distributed network. Not only is it important to have better distributed points overall, it is particularly important to promote measurements within underrepresented biomes. Having better representation of these biomes can illuminate patterns and responses not currently well described. Increased field measurements of Rh are also paramount to increasing our understanding of global Rs dynamics. Not only does this benefit bottom-up processes using machine learning, but also process-based models and other estimation methods. Clearly, we as a community need to decrease uncertainty, but the question remains as to which sources are driving the majority of uncertainty. That is, whether the spatial bias of the Rs observational data is driving the majority of uncertainty, or if it is due to other factors such as sub-pixel

heterogeneity or uncertainty in covariate data layers. However, these potential sources of error cannot be separated without additional measurements in targeted locations. Constraining the global Rs flux will require further intercomparisons and collaborations between scientists monitoring in situ Rs, the remote sensing community, and process-based modelers. This underscores the need for transparency, collaboration, and data sharing in environmental research.

Chapter 3

CONCLUSIONS AND MOVING FORWARD

We modeled and mapped global Rs at a 1km resolution and used empirical equations to derive global Rh estimates. We additionally investigated the mean and sum Rs within individual land cover classifications. Further, we identified optimal future sampling locations around the globe to facilitate better constrained global Rs estimates, including those within under-represented biomes. These findings highlight the need for increased global collaboration within the global carbon cycling community.

Moving forward, it is important to reduce uncertainty in global Rs models. Our proof of concept models showed that a better distributed Rs network would reduce global Rs uncertainty. Not only is it important to have better distributed points overall, it is particularly important to promote measurements within underrepresented biomes. Having better representation of these biomes can illuminate patterns and responses not currently well described. Increased field measurements of Rh are also paramount to increasing our understanding of global Rs dynamics. Not only does this benefit bottom-up processes using machine learning, but also process-based models and other estimation methods. Clearly, growing databases with new and updated measurements is important for increased understanding of global processes, but it is not enough to say that we simply need more data- we need increased data in targeted locations. Not only do we need to decrease uncertainty in Rs models, we need to decrease the bias in Rs measurement locations. These measurements are used for benchmarking of earth

system models, and biased measurement locations leads to biased benchmarking, affecting the overall models. Even with additional measurements in underrepresented areas, the question remains as to whether the bias towards northern temperate locations can be significantly reduced, considering the wealth of measurements already located in these regions. Nevertheless, moving forward, we as a community need more targeted measurements taken in underrepresented regions.

There still exists many knowledge gaps in modeling Rs at the regional to global scale. Many global Rs studies are at an annual timescale, but the creation of three additional publicly accessible global Rs databases since 2018 offers many opportunities for analyzing global Rs at finer temporal scales. It is important to have both finer spatial and temporal resolutions to improve our understanding of the links and pattens between local to global Rs rates and estimates. Working at finer spatial and temporal scales is possible with high performance computing resources and increases in available data. Utilizing MAT within inaccessible regions to estimate Rs in combination with meteorological data and current measurement sites could be used to improve global Rs models (Jian et al. 2020b). As the field of data science further evolves, there exist opportunities to model global Rs via deep learning, a subcategory of machine learning that involves large neural networks that has been successfully used in other fields, and other state of the art techniques.

This line of research has implications for land carbon uptake, quantifying the carbon budget, climate system trends, land management, policy on protecting soils,

quantify balance between Rs and carbon uptake (see carbon cycle diagram), and strategies for implementing optimized sampling designs.

REFERENCES

Bahn M, Reichstein M, Davidson EA, et al (2010) Soil respiration at mean annual temperature predicts annual total across vegetation types and biomes. Biogeosciences 7:2147–2157. doi: 10.5194/bg-7-2147-2010

Basu, S., Kumbier, K., Brown, J. B., & Yu, B. (2018). Iterative random forests to discover predictive and stable high-order interactions. *Proceedings of the National Academy of Sciences*, *115*(8), 1943–1948. https://doi.org/10.1073/pnas.1711236115

Batjes N.H. (1996) Total carbon and nitrogen in the soils of the world. European Journal of Soil Science 47:151–163. doi: 10.1111/j.1365-2389.1996.tb01386.x

Beer C, Reichstein M, Tomelleri E, et al (2010) Terrestrial Gross Carbon Dioxide Uptake: Global Distribution and Covariation with Climate. Science 1184984. doi: 10.1126/science.1184984

Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST*, *25*(2), 197–227. https://doi.org/10.1007/s11749-016-0481-7

Bond-Lamberty B, Wang C, Gower ST (2004) A global relationship between the heterotrophic and autotrophic components of soil respiration? Global Change Biology 10:1756–1766. doi: 10.1111/j.1365-2486.2004.00816.x

Bond-Lamberty B, Thomson A (2010) A global database of soil respiration data. Biogeosciences 7:1915–1926. doi: 10.5194/bg-7-1915-2010

Bond-Lamberty B (2018) New Techniques and Data for Understanding the Global Soil Respiration Flux. Earth's Future 6:1176–1180. doi: 10.1029/2018EF000866

Bond-Lamberty B, Bailey VL, Chen M, et al (2018) Globally rising soil heterotrophic respiration over recent decades. Nature 560:80. doi: 10.1038/s41586-018-0358-x

Bond-Lamberty, B., Christianson, D. S., Malhotra, A., Pennington, S. C., Sihi, D., AghaKouchak, A., Anjileli, H., Arain, M. A., Armesto, J. J., Ashraf, S., Ataka, M., Baldocchi, D., Black, T. A., Buchmann, N., Carbone, M. S., Chang, S.-C., Crill, P., Curtis, P. S., Davidson, E. A., … Zou, J. (2020). COSORE: A community database for continuous soil respiration and other soil-atmosphere greenhouse gas flux data. *Global Change Biology*. https://doi.org/10.1111/gcb.15353

Breiman, L. (2001) Random forests. *Machine Learning*, 45(1), 5-32. doi:10.1023/A:1010933404324.

Channan, S., Collins, K., & Emanuel, W.R. (2014). Global mosaics of the standard MODIS land cover type data (Vol. 30). College Park, Maryland, USA: University of Maryland and the Pacific Northwest National Laboratory.

Chapin, F. S., Woodwell, G. M., Randerson, J. T., Rastetter, E. B., Lovett, G. M., Baldocchi, D. D., Clark, D. A., Harmon, M. E., Schimel, D. S., Valentini, R., Wirth, C., Aber, J. D., Cole, J. J., Goulden, M. L., Harden, J. W., Heimann, M., Howarth, R. W., Matson, P. A., McGuire, A. D., … Schulze, E.-D. (2006). Reconciling Carbon-cycle Concepts, Terminology, and Methods. *Ecosystems*, *9*(7), 1041–1050. https://doi.org/10.1007/s10021-005-0105-7

Chen S, Huang Y, Zou J, et al (2010) Modeling interannual variability of global soil respiration from climate and soil properties. Agricultural and Forest Meteorology 150:590–605. doi: 10.1016/j.agrformet.2010.02.004

Chen S, Huang Y, Xie W, et al (2013) A new estimate of global soil respiration from 1970 to 2008. Chinese Science Bulletin 58:4153–4160. doi: 10.1007/s11434-013-5912-1

Chen S, Zou J, Hu Z, et al (2014) Global annual soil respiration in relation to climate, soil properties and vegetation characteristics: Summary of available data. Agricultural and Forest Meteorology 198–199:335–346. doi: 10.1016/j.agrformet.2014.08.020

Chu, H.-J., Lin, Y.-P., Jang, C.-S., & Chang, T.-K. (2010). Delineating the hazard zone of multiple soil pollutants by multivariate indicator kriging and conditioned Latin hypercube sampling. *Geoderma*, *158*(3), 242–251. https://doi.org/10.1016/j.geoderma.2010.05.003

Contreras, J., Ballari, D., Bruin, S. de, & Samaniego, E. (2019). Rainfall monitoring network design using conditioned Latin hypercube sampling and satellite precipitation estimates: An application in the ungauged Ecuadorian Amazon. *International Journal of Climatology*, *39*(4), 2209–2226. https://doi.org/10.1002/joc.5946

Davidson EA, Janssens IA, Luo Y (2006) On the variability of respiration in terrestrial ecosystems: moving beyond Q10. Global Change Biology 12:154–164. doi: 10.1111/j.1365-2486.2005.01065.x

Enting, I. G., Rayner, P. J., & Ciais, P. (2012). Carbon Cycle Uncertainty in REgional Carbon Cycle Assessment and Processes (RECCAP). *Biogeosciences*, *9*(8), 2889–2904. https://doi.org/10.5194/bg-9-2889-2012

Epule, T. E. (2015). A New Compendium of Soil Respiration Data for Africa. *Challenges*, *6*(1), 88–97. https://doi.org/10.3390/challe6010088

Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, *37*(12), 4302–4315. https://doi.org/10.1002/joc.5086

Friedlingstein P, Cox P, Betts R, et al (2006) Climate–Carbon Cycle Feedback Analysis: Results from the C4MIP Model Intercomparison. J Climate 19:3337–3353. doi: 10.1175/JCLI3800.1

Friedlingstein P, Meinshausen M, Arora VK, et al (2014) Uncertainties in CMIP5 Climate Projections due to Carbon Cycle Feedbacks. J Climate 27:511–526. doi: 10.1175/JCLI-D-12-00579.1

Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, *31*(14), 2225–2236. https://doi.org/10.1016/j.patrec.2010.03.014

Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2015). VSURF: An R Package for Variable Selection Using Random Forests. *The R Journal* 7, 15:19-33.

Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E., (2015) Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation. Journal of Computational and Graphical Statistics 24:1, 44-65, DOI: 10.1080/10618600.2014.907095

Greenwell, B., M., (2017) pdp: An R Package for Constructing Partial Dependence Plots. The R Journal 9, 421-436. https://doi.org/10.32614/RJ-2017-016

Hashimoto S (2012) A New Estimation of Global Soil Greenhouse Gas Fluxes Using a Simple Data-Oriented Model. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0041962.

Hashimoto S, Carvalhais N, Ito A, et al (2015) Global spatiotemporal distribution of soil respiration modeled using a global database. Biogeosciences Discussions 12:4331–4364. doi: 10.5194/bgd-12-4331-2015

Hibbard KA, Law BE, Reichstein M, Sulzman J (2005) An analysis of soil respiration

across northern hemisphere temperate ecosystems. Biogeochemistry 73:29–70. doi: 10.1007/s10533-004-2946-0

Hoffman, F. M., Koven, C. D., Keppel-Aleks, G., Lawrence, D. M., Riley, W. J., Randerson, J. T., Ahlström, A., Abramowitz, G., Baldocchi, D. D., Best, M. J., Bond-Lamberty, B., De Kauwe, M. G., Denning, A. S., Desai, A. R., Eyring, V., Fisher, J. B., Fisher, R. A., Gleckler, P. J., Huang, M., … Koch, D. (2017). *2016 International Land Model Benchmarking (ILAMB) Workshop Report* (DOE/SC-0186). USDOE Office of Science, Washington, DC (United States). https://doi.org/10.2172/1330803

Hursh A, Ballantyne A, Cooper L, et al (2017) The sensitivity of soil respiration to soil temperature, moisture, and carbon supply at the global scale. Global Change Biology 23:2090–2103. doi: 10.1111/gcb.13489

Jian J, Steele MK, Thomas RQ, et al (2018a) Constraining estimates of global soil respiration by quantifying sources of variability. Global Change Biology 24:4143–4159. doi: 10.1111/gcb.14301

Jian J, Steele MK, Day SD, Thomas RQ (2018b) Future global soil respiration rates will swell despite regional decreases in temperature sensitivity caused by rising temperature. Earth's Future, 6. doi: 10.1029/2018EF000937

Jian J, Steele MK, Day SD, et al (2018c) Measurement strategies to account for soil respiration temporal heterogeneity across diverse regions. Soil Biology and Biochemistry 125:167–177. doi: 10.1016/j.soilbio.2018.07.003

Jian, J., Vargas, R., Anderson-Teixeira, K., Stell, E., Herrmann, V., Horn, M., Kholod, N., Manzon, J., Marchesi, R., Paredes, D., and Bond-Lamberty, B. (in review, 2020a) A restructured and updated global soil respiration database (SRDB-v5), *Earth Syst. Sci. Data Discuss.*, https://doi.org/10.5194/essd-2020-136.

Jian, J., Bahn, M., Wang, C., Bailey, V. L., & Bond-Lamberty, B. (2020b). Prediction of annual soil respiration from its flux at mean annual temperature. *Agricultural and Forest Meteorology*, *287*, 107961. https://doi.org/10.1016/j.agrformet.2020.107961

Lin, Y.-P., Chu, H.-J., Wang, C.-L., Yu, H.-H., & Wang, Y.-C. (2009). Remote Sensing Data with the Conditional Latin Hypercube Sampling and Geostatistical Approach to Delineate Landscape Changes Induced by Large Chronological Physical Disturbances. *Sensors*, *9*(1), 148–174. https://doi.org/10.3390/s90100148

Malhi, Y., Grace, J. (2000). Tropical forests and atmospheric carbon dioxide. *Trends in Ecology & Evolution*, *15*(8), 332–337. https://doi.org/10.1016/S0169-5347(00)01906-6

Meinshausen, N. (2006). Quantile Regression Forests. *Journal of Machine Learning Research, 7*, 983-999.

Minasny, B., & McBratney, A. B. (2006). A conditioned Latin hypercube method for sampling in the presence of ancillary information. *Computers & Geosciences*, *32*(9), 1378–1388. https://doi.org/10.1016/j.cageo.2005.12.009

Mulder, V. L., de Bruin, S., & Schaepman, M. E. (2013). Representing major soil variability at regional scale by constrained Latin Hypercube Sampling of remote sensing data. *International Journal of Applied Earth Observation and Geoinformation*, *21*, 301–310. https://doi.org/10.1016/j.jag.2012.07.004

Poulter, B., Frank, D., Ciais, P., Myneni, R. B., Andela, N., Bi, J., Broquet, G., Canadell, J. G., Chevallier, F., Liu, Y. Y., Running, S. W., Sitch, S., & van der Werf, G. R. (2014). Contribution of semi-arid ecosystems to interannual variability of the global carbon cycle. *Nature*, *509*(7502), 600–603. https://doi.org/10.1038/nature13376

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Raich JW, Potter CS, Bhagawati D (2002) Interannual variability in global soil respiration, 1980–94. Global Change Biology 8:800–812. doi: 10.1046/j.1365-2486.2002.00511.x

Raich JW, Potter CS (1995) Global patterns of carbon dioxide emissions from soils. Global Biogeochemical Cycles 9:23–36. doi: 10.1029/94GB02723

Raich JW, Schlesinger WH (1992) The global carbon dioxide flux in soil respiration and its relationship to vegetation and climate. Tellus B 44:81–99. Doi: 10.1034/j.1600-0889.1992.t01-1-00001.x

Raich JW, Tufekcioglu A (2000) Vegetation and soil respiration: Correlations and controls. Biogeochemistry 48:71-90. Doi: https://doi.org/10.1023/A:1006112000616

Reichstein M, Rey A, Freibauer A, et al (2003) Modeling temporal and large-scale

spatial variability of soil respiration from soil water availability, temperature and vegetation productivity indices. Global Biogeochemical Cycles 17:. doi: 10.1029/2003GB002035

Reichstein M, Beer C (2008) Soil respiration across scales: The importance of a model–data integration framework for data interpretation. Journal of Plant Nutrition and Soil Science 171:344–354. doi: 10.1002/jpln.200700075

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195. https://doi.org/10.1038/s41586-019-0912-1

Reuter, H.I., & Hengl, T. (2016) Worldgrids- A Public Repository of Global Soil Covariates. *Computing Ethics: A Multicultural Approach* (pp. 287-291). Boca Raton, FL., USa: CRC Press.

Ryan MG, Law BE (2005) Interpreting, measuring, and modeling soil respiration. *Biogeochemistry* 73:3–27. doi: 10.1007/s10533-004-5167-7

Savage KE, Davidson EA (2003) A comparison of manual and automated systems for soil CO2 flux measurements: trade-offs between spatial and temporal resolution. *J Exp Bot* 54:891–899. doi: 10.1093/jxb/erg121

Schlesinger WH (1977) Carbon Balance in Terrestrial Detritus. Annual Review of Ecology and Systematics 8:51–81. doi: 10.1146/annurev.es.08.110177.000411

Schlesinger, W. H., & Andrews, J. A. (2000). Soil respiration and the global carbon cycle. *Biogeochemistry*, *48*(1), 7–20. https://doi.org/10.1023/A:1006247623877

Shao, P., Zeng, X., Moore, D. J. P., & Zeng, X. (2013). Soil microbial respiration from observations and Earth System Models. *Environmental Research Letters*, *8*(3), 034034. https://doi.org/10.1088/1748-9326/8/3/034034

Silva, S. H. G., Owens, P. R., Silva, B. M., Oliveira, G. C. de, Menezes, M. D. de, Pinto, L. C., & Curi, N. (2015). Evaluation of Conditioned Latin Hypercube Sampling as a Support for Soil Mapping and Spatial Variability of Soil Properties. *Soil Science Society of America Journal*, *79*(2), 603–611. https://doi.org/10.2136/sssaj2014.07.0299

Singh JS, Gupta SR (1977) Plant Decomposition and Soil Respiration in Terrestrial

Ecosystems. Botanical Review 43:449–528. https://www.jstor.org/stable/4353928?seq=1#page_scan_tab_contents

Smith, F. B., & Brown, P. E. (1931). Soil Respiration 1. *Agronomy Journal*, *23*(11), 909–916. https://doi.org/10.2134/agronj1931.00021962002300110005x

Stumpf, F., Schmidt, K., Behrens, T., Schönbrodt-Stitt, S., Buzzo, G., Dumperth, C., Wadoux, A., Xiang, W., & Scholten, T. (2016). Incorporating limited field operability and legacy soil samples in a hypercube sampling design for digital soil mapping. *Journal of Plant Nutrition and Soil Science*, *179*(4), 499–509. https://doi.org/10.1002/jpln.201500313

Subke, J.-A., Inglima, I., & Cotrufo, M. F. (2006). Trends and methodological impacts in soil CO2 efflux partitioning: A metaanalytical review. *Global Change Biology*, *12*(6), 921–943. https://doi.org/10.1111/j.1365-2486.2006.01117.x

Tang, X., Fan, S., Du, M., Zhang, W., Gao, S., Liu, S., Chen, G., Yu, Z., & Yang, W. (2020). Spatial and temporal patterns of global soil heterotrophic respiration in terrestrial ecosystems. *Earth System Science Data*, *12*(2), 1037–1051. https://doi.org/10.5194/essd-12-1037-2020

Thessen, A. (2016). Adoption of Machine Learning Techniques in Ecology and Earth Science. *One Ecosystem*, *1*, e8621. https://doi.org/10.3897/oneeco.1.e8621

Vaysse, K., & Lagacherie, P. (2017). Using quantile regression forest to estimate uncertainty of digital soil mapping products. *Geoderma*, *291*, 55–64. https://doi.org/10.1016/j.geoderma.2016.12.017

Vargas R, Carbone MS, Reichstein M, Baldocchi DD (2011) Frontiers and challenges in soil respiration research: from measurements to model-data integration. *Biogeochemistry* 102:1–13. doi: 10.1007/s10533-010-9462-1

Villarreal, S., Guevara, M., Alcaraz-Segura, D., & Vargas, R. (2019). Optimizing an Environmental Observatory Network Design Using Publicly Available Data. *Journal of Geophysical Research: Biogeosciences*, *124*(7), 1812–1826. https://doi.org/10.1029/2018JG004714

Warner, D. L., Bond-Lamberty, B., Jian, J., Stell, E., & Vargas, R. (2019). Spatial Predictions and Associated Uncertainty of Annual Soil Respiration at the Global Scale. *Global Biogeochemical Cycles*, *33*(12), 1733–1745. https://doi.org/10.1029/2019GB006264

Xu M, Shang H (2016) Contribution of soil respiration to the global carbon equation.

Journal of Plant Physiology 203:16–28. doi: 10.1016/j.jplph.2016.08.007

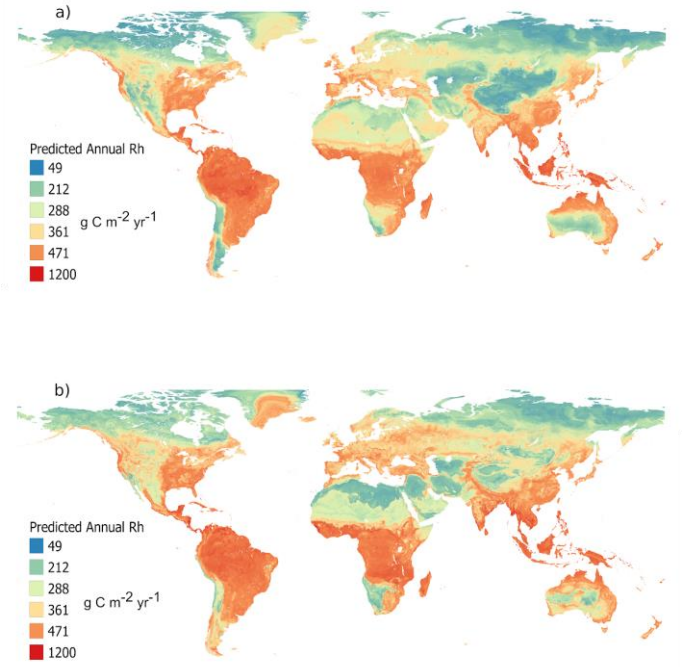
Yin, G., Li, A., Zeng, Y., Xu, B., Zhao, W., Nan, X., Jin, H., & Bian, J. (2016). A Cost-Constrained Sampling Strategy in Support of LAI Product Validation in Mountainous Areas. *Remote Sensing*, *8*(9), 704. https://doi.org/10.3390/rs8090704

Yin, G., Li, A., & Verger, A. (2017). Spatiotemporally Representative and Cost-Efficient Sampling Design for Validation Activities in Wanglang Experimental Site. *Remote Sensing*, *9*(12), 1217. https://doi.org/10.3390/rs9121217

Zhao Z, Peng C, Yang Q, et al (2017) Model prediction of biome-specific global soil respiration from 1960 to 2012. Earth's Future 5:715–729. doi: 10.1002/2016EF000480

Appendix A

SUPPLEMENTARY MATERIAL



Visualization of v3 (a) and v5 (b) mean Rh calculated using the Bond-Lamberty 2004 method. Units are in g C m^{-2} yr^{-1}.