ASIAN JOURNAL OF FORESTRY
Volume 5, Number 2, December 2021 E-ISSN: 2580-2844
Pages: 111-122 DOI: 10.13057/asianjfor/r050207

Filtering multi-collinear predictor variables from multi-resolution rasters of WorldClim 2.1 for Ecological Niche Modeling in Indonesian context

PRAKASH PRADHAN[1,♥], AHMAD DWI SETYAWAN[2,3]

[1]West Bengal Biodiversity Board, Prani Sampad Bhawan, 5th Floor, LB-2, Salt Lake, Sector-III, Kolkata, PIN – 700106, West Bengal, India.
♥email: shresthambj@gmail.com

[2]Department of Environmental Science, Faculty of Mathematics and Natural Sciences, Universitas Sebelas Maret. Jl. Jend. Urip Sumoharjo No. 179, Surakarta 57128, Central Java, Indonesia

[3]Department of Biology, Faculty of Mathematics and Natural Sciences, Universitas Sebelas Maret. Jl. Ir. Sutami 36A, Surakarta 57126, Central Java, Indonesia

Manuscript received: 8 August 2021. Revision accepted: 16 September 2021.

Abstract. *Pradhan P, Setyawan AD. 2021. Filtering multi-collinear predictor variables from multi-resolution rasters of WorldClim 2.1 for Ecological Niche Modeling in Indonesian context. Asian J For 5: 111-122.* WorldClim is one of the popular environmental datasets which hosts multi-resolution interpolated gridded climate raster surfaces and derived bioclimatic variables for both the immediate past, present and future scenarios. Bioclimatic variables along with other environmental factors like solar radiation, wind speed, water vapour pressure etc. have been used as primary set of explanatory variables for mapping and spatial modeling of many biological processes, including defining environmental niche of a species and identifying potential areas for its distribution through machine learning methods like Ecological Niche Modeling or Species Distribution Modeling or Habitat Suitability Modeling. However, the interpolated explanatory datasets are known to cause over-fitting of the models mainly due to multi-collinearity or redundancy within the variables. In the present study, 58 bioclimatic and environmental variables of Indonesian extent extracted from WorldClim 2.1 are screened to investigate the presence of multi-collinearity or redundancy. From the total 3364 variable pairs per raster resolution, 174 variable pairs were known to be affected by multicollinearity, from which temperature related bioclimatic variables, water vapour pressure and elevation associated variables were highly notable. For all the raster resolutions, bioclimatic variable 2, 3, 4, 15, 18 and 19, as well as slope, aspect, solar radiation for January, April, May, September, wind speed for August and November were found to be non-collinear. While, solar radiation for March and July were found to be non-collinear for 30s, 2.5m and 5m raster resolutions; Wind speed of July was non-collinear for 30s and 2.5m; Solar radiation for February and June were non-collinear for 10m; water vapour pressure for August for 2.5m and wind speed for January was non-collinear for 30s raster resolutions. The results of this study might serve as a convenient reference for investigators of the region for selection of bioclimatic and other environmental variables for conducting ecological niche modeling studies.

Keywords: Bioclimatic variables, elevation, habitat suitability modeling, MaxEnt, R, raster resolution, solar radiation, species distribution modeling, variance inflation factor, water vapour pressure, wind speed

INTRODUCTION

Bioclimatic variables along with other environmental variables represent important explanatory/predictor variables to understand species distribution (Busby 1986; Nix 1986). Bioclimatic variables express spatial variation in annual means, seasonality and extreme or limiting climatic factors, and represent biologically meaningful parameters for characterizing species distributions (Saatchi et al. 2008; O'Donnell and Ignizio 2012; Pradhan 2015). The advent of machine learning based ecological niche modeling (ENM)/species distribution modeling (SDM)/habitat suitability modeling (HSM) – from now on in this paper it is called ecological niche modeling – has opened an array of utility of bioclimatic variables and other climatic surfaces.

Among several climate and environmental databases, WorldClim is one of the popular environmental datasets used for mapping and spatial modeling of many biological processes due to availability of multi-resolution interpolated gridded climate raster surfaces and derived bioclimatic variables for both the immediate past, present and future scenarios. WorldClim version 1.4 was developed by Hijmans et al. (2005) with the present 'year 1960-1990' and future climate surfaces based on Coupled Model Intercomparison Project Phase 5 (CMIP5). The updated and expanded WorldClim version 2 was released during 2017 by Fick and Hijmans (2017) and it was further upgraded to version 2.1 (released on January 2020) with the present 'year 1970-2000' and future climate surfaces based on CMIP6, as well as monthly environmental variables such as solar radiation, wind speed and water vapour pressure. However, one of the major drawback linked to such interpolation derived environmental datasets is reported to be redundancy or multicollinearity and overfitting of resultant models (Pradhan 2016).

In machine learning process, optimal training by minimal set of explanatory variables (low training error) is very important for building optimal model, which could perform well against testing variables (low testing error).

However, training based on redundant, multicollinear, more than necessary and less relevant inputs may lead to learning of 'noise' of training data to the resultant model making it fit close to the training data (overfitted/low training error) and make it more complex and less sensitive to testing data (high testing error) (Anderson and Gonzales 2011; van Gils et al. 2014).

However, there are methods to counter such anomaly by selecting few non-collinear explanatory/predictor variables, making the resultant models less-overfitted and are simpler and parsimonius/based on minimally selected predictor variables. Variance Inflation Factor (VIF) analysis is a widely used method to identify problematic collinearity/redundancy among the variables. VIF is an indicator of the degree to which the standard errors are inflated due to the levels of multicollinearity (Montgomery and Peck 1992). In R, VIF can be calculated in packages like *car* (Fox and Weisberg 2019), *faraway* (Faraway 2016), *usdm* (Naimi et al. 2014), *vegan* (Oksanen et al. 2016), etc. However, these packages provide only individual VIFs per variable when the VIF itself being a derivative of correlation, pair-wise calculation of VIF would have provided more insight into which two variables are collinear and at which level of VIF.

Ecological niche modeling has been increasingly employed in Indonesia for modeling present habitat suitability and in some cases future potential distribution. Some notable studies from the region employing ENM techniques using WorldClim data are: Proboscis Monkey (*Nasalis larvatus*) (Suwarto et al. 2016), Javan hawk-eagle *Nisaetus bartelsi* (Nursamsi et al. 2018), Zebra Wood (*Guettarda speciosa*) (Yudaputra et al. 2019), *Baccaurea macrocarpa* trees (Gunawan et al. 2021), *Selaginella* spp. (Setyawan et al. 2017, 2020a, 2020b, 2021) and so on. While such studies provide strong basis for the development of ecological niche modeling application in Indonesia, methodological improvements to minimize the presence of multicollinearity might be useful for future application of ENM.

The current analysis aimed to investigate the presence of multicollinearity among 19 bioclimatic variables and monthly environmental variables of solar radiation, wind speed and water vapour pressure available in WorldClim version 2.1 in Indonesian context. In doing so, we undertake from the scratch approach in R for studying pair-wise multicollinearity in terms of Variance Inflation Factor amongst such variables. This analysis suggests variable pairs which are not to be used in combination together and also to identify the purely non-collinear variables which may be used for easy reference for future studies applying ecological niche modeling in the region.

MATERIALS AND METHODS

Data requirements

WorldClim version 2.1 website hosts multi-resolution continuous raster surfaces of monthly climate data of minimum, mean, and maximum temperature (°C), precipitation (mm), solar radiation (kJ m^{-2} day^{-1}), wind speed (m s^{-1}), water vapour pressure (kPa) for the 'current/present' period of 1970-2000 in *.tiff (Geotiff) format, besides hosting rasters of 19 bioclimatic variables for the above-mentioned period. For the current study, rasters of bioclimatic variables, solar radiation, wind speed, water vapour pressure were downloaded from WorldClim version 2.1 website (Fick and Hijmans 2017) in four raster resolutions viz. 30 seconds, 2.5 minutes, 5 minutes and 10 minutes. Raster resolution here is referred to the cell size/ spatial resolution which is the dimension of the area covered on the map and represented by a single pixel.

Raster handling workflow in R

By default, getData function of *raster* package still downloads obsolete WorldClim version 1.4. Hence, the updated 2.1 version raster files were downloaded, sorted resolution-wise, loaded in R, then resolution-wise raster *stack* was prepared with *raster* (Hijmans 2020) and *rgdal* (Bivand et al. 2020) packages, and projected to WGS84 coordinate reference system. Rasters of slope and aspect were created from Digital Elevation Model (DEM) with *terrain* function of *raster* package. Shape file of the world in 1:1 million scale was downloaded from EUROSTAT (2020) and the rasters were cropped and masked to the extent of Indonesia. Rasters were renamed as Bio1-19 for bioclimatic variables, Elev for DEM raster, Srad1-12 respectively for twelve monthly solar radiation rasters, Vapr1-12 respectively for twelve monthly water vapour pressure rasters, Wind1-12 respectively for monthly wind speed rasters, while names of slope and aspect rasters were kept unchanged.

Statistical analysis

The cropped rasters were converted to matrix through *as.matrix* function, and to obtain Pearson r value, pair-wise correlation of matrices (N) were conducted through *cor* function. After getting r value of the matrix, r^2 was obtained by multiplying N with N. Pair-wise Variance Inflation Factor (VIF) was calculated through the formula $VIF<-(1/(1-r^2))$. After VIF calculation, all cells with *Inf* values were replaced with 1 and the table values were formatted to one digit after zero through *round* function. Further, the VIF matrix was converted to dataframe, eliminating diagonal and duplicates to get pair-wise VIF. Shapiro test for assessing normality of data, Kruskal-Wallis and Friedman rank sum test, as well as post hoc Pair-wise Wilcox test were carried out in R with *shapiro.test()*, *kruskal.test()*, *friedman.test()* and *pairwise.wilcox.test()* functions respectively.

Data visualization

Density plot and Boxplot were prepared through *ggplot2* library. Heatmap of the pair-wise VIF matrix (used only for comparison of result of *usdm* package) was created with *heatmaply* package (Galili 2017), the resultant plot was saved as webpage through *htmlwidgets* package (Vaidyanathan 2019), and the webpage was converted to jpeg through *webshot2* package (Chang 2020). For visual comparison of four raster cell sizes with reference to zoomed inset of Gunung Leuser National Park region,

northern Sumatra, cropped and masked rasters of four resolutions of mean temperature of warmest quarter (Bio10) were visualized in R with plot function and exported as *.pdf. Four individual maps were compiled in Adobe Photoshop.

RESULTS AND DISCUSSION

In this study, gridded rasters of 58 explanatory variables available at WorldClim 2.1 were analyzed for presence of multicollinearity among them at the masked extent of Indonesia. The studied variables include 19 bioclimatic variables, elevation and its derivatives of slope and aspect, monthly variables of solar radiation (Srad), wind speed and water vapour pressure (Vapr). It was found that 174 pairs of above environmental variables had multicollinear relationship in one raster resolution or the other. The highest VIF for all raster resolutions belonged to the variable pair Bio1-Bio10 263±51.2

The environmental variables which showed multicollinearity were Bio1 (annual mean temperature), Bio5 (max temperature of warmest month), Bio6 (min temperature of coldest month), Bio8 (mean temperature of wettest quarter), Bio9 (mean temperature of driest quarter), Bio10 (mean temperature of warmest quarter), Bio11 (mean temperature of coldest quarter), Bio13 (precipitation of wettest month), Bio14 (precipitation of driest month), Bio16 (precipitation of wettest quarter), Bio17 (precipitation of driest quarter), elev (elevation/altitude), solar radiation for October (Srad10) and November (Srad11), water vapour pressure for January-December (Vapr1-12), wind for January-July (Wind1-Wind7), wind for September, October and December (Wind9, Wind10 and Wind 12). The detailed pair-wise comparison of multicollinear environmental variables are discussed below.

VIF pairs corresponding to Bio 1 (Annual Mean Temperature)

Bio 1 or annual mean temperature approximates the total energy inputs for an ecosystem. Results suggest that for four raster resolutions (30s, 2.5m, 5m and 10m), Bio1 is not to be used along with Bio10, Bio11, Bio5, Bio6, Bio8, Bio9, Elev, Vapr1, Vapr2, Vapr3, Vapr4, Vapr5, Vapr11, Vapr12. In addition, for 2.5m, 5m and 10m resolution rasters, Bio1 is not to be used along with Vapr10, and for 10m resolution raster, Bio1 is to be avoided to be used together with Vapr6.

VIF pairs corresponding to Bio 5 (Maximum Temperature of the Warmest Month)

Bio5 or maximum temperature of warmest month indicates the maximum monthly temperature occurrence over a given year (time-series) or averaged span of years (normal). This information is useful when examining whether species distributions are affected by warm temperature anomalies throughout the year. Results suggest that for four raster resolutions (30s, 2.5m, 5m and 10m) rasters, Bio5 is not to be used along with Bio1 (Table 1), Bio10, Bio11, Bio6, Bio8, Bio9, Elev, Vapr1, Vapr2, Vapr3, Vapr12 (Table 2). For 5m and 10m resolution rasters, Bio5 is not to be used along with Vapr11, while for 10m resolution raster, Bio5 is not to be used along with Vapr5.

VIF pairs corresponding to Bio 6 (Minimum Temperature of the Coldest Month)

Bio 6 or minimum temperature of coldest month takes account of the minimum temperature value across all months within a given year. This index is useful when examining whether species distributions are affected by cold temperature anomalies throughout the year. Results suggest that for four raster resolutions (30s, 2.5m, 5m and 10m), Bio6 is not to be used along with Bio10, Bio11, Bio8, Bio9, Elev, Vapr1, Vapr2, Vapr3, Vapr4, Vapr5, Vapr6, Vapr10, Vapr11, Vapr12 (Table 3), Bio1 (Table 1), Bio5 (Table 2). For 10m resolution raster, Bio6 is not to be used along with Vapr7, Vapr8, Vapr9.

Table 1. Distribution of VIF in variable pair of Bio1 (annual mean temperature) with other variables corresponding to 30 arc second, 2.5 minutes, 5 minutes and 10 minutes raster resolution. VIF values with (*) indicate values <10

VarPair	30s	2.5m	5m	10m
Bio1-Bio10	212.9	237.3	271.6	330.9
Bio1-Bio11	94.5	102.6	112.6	130.2
Bio1-Bio5	40.8	45.5	52.2	64
Bio1-Bio6	41.8	46.5	51.3	58.3
Bio1-Bio8	77.4	83.1	88.6	97.9
Bio1-Bio9	55.3	60.3	65.8	75
Bio1-Elev	28.9	30.5	31.4	32.2
Bio1-Vapr1	17.2	19.5	21.8	26.4
Bio1-Vapr10	9.5*	10.3	11.1	12.7
Bio1-Vapr11	13.3	14.3	15.3	17.2
Bio1-Vapr12	16.2	18.1	20.1	23.8
Bio1- Vapr2	17.7	20.1	22.7	27.2
Bio1-Vapr3	13.6	15	16.5	19.1
Bio1-Vapr4	12.6	13.9	15.3	17.9
Bio1-Vapr5	11.1	12.2	13.4	15.8
Bio1-Vapr6	8*	8.7*	9.6*	11.3

Table 2. Distribution of VIF in variable pair of Bio5 (maximum temperature of warmest month) with other variables corresponding to 30 arc second, 2.5 minutes, 5 minutes and 10 minutes raster resolution. VIF values with (*) indicate values <10. Reverse variable pairs of Bio5 available in other tables are not mentioned here.

VarPair	30s	2.5m	5m	10m
Bio5-Bio10	57.3	64	71.3	84.2
Bio5-Bio11	22.2	24.4	27.8	33.6
Bio5-Bio6	11.8	13.1	14.8	17.4
Bio5-Bio8	24.7	26.7	29.4	34.2
Bio5-Bio9	22.1	24.8	28.5	34.6
Bio5-Elev	14.9	16.2	17.4	19.1
Bio5- Vapr1	12.1	13.9	15.8	19.6
Bio5-Vapr12	10.7	12.1	13.6	16.6
Bio5-Vapr2	12.5	14.5	16.6	20.5
Bio5-Vapr3	10.3	11.6	13.1	15.9
Bio5-Vapr11	8.5*	9.4*	10.3	12.1
Bio5-Vapr4	8.6*	9.8*	11	13.4
Bio5-Vapr5	7*	7.8*	8.9*	10.9

VIF pairs corresponding to Bio 8 (Mean Temperature of the Wettest Quarter)

The quarterly index of Bio 8 or mean temperature of wettest quarter approximates average temperature for the three months with the highest cumulative precipitation. This index provides mean temperatures during the wettest three months of the year, which can be useful for examining how such environmental factors may affect species seasonal distributions.

Results suggest that for four raster resolutions (30s, 2.5m, 5m and 10m), Bio8 is not to be used along with Bio10, Bio11, Bio9, Elev, Vapr1, Vapr2, Vapr3, Vapr11, Vapr12 (Table 4), Bio1 (Table 1), Bio5 (Table 2), Bio6 (Table 3). For 5m and 10m resolution rasters, Bio8 is not to be used along with Vapr4, while for 10m resolution raster, Bio8 is not to be used along with Vapr5.

VIF pairs corresponding to Bio 9 (Mean Temperature of the Driest Quarter)

This quarterly index of Bio 9 or mean temperature of driest quarter approximates mean temperature for three months of the year with the lowest cumulative precipitation, which can be useful for examining how such environmental factors may affect species seasonal distributions. Results suggest that for four raster resolutions (30s, 2.5m, 5m and 10m), Bio9 is not to be used along with Bio10, Bio11, Elev, Vapr1, Vapr2, Vapr3, Vapr4, Vapr5, Vapr10, Vapr11, Vapr12 (Table 5), Bio1 (Table 1), Bio5 (Table 2), Bio6 (Table 3), Bio8 (Table 4). In addition, for 10m resolution raster, Bio9 is not to be used along with Vapr6.

VIF pairs corresponding to Bio 10 (Mean Temperature of the Warmest Quarter)

The quarterly index of Bio10 or mean temperature of warmest quarter approximates mean temperatures that prevail during the warmest three months of the year, which can be useful for examining how such environmental factors may affect species seasonal distributions. Results suggest that for four raster resolutions (30s, 2.5m, 5m and 10m), Bio 10 or mean temperature of warmest quarter is not to be used along with Bio11, Elev, Vapr1, Vapr2, Vapr3, Vapr4, Vapr11, Vapr12, Bio1 (Table 1), Bio5 (Table 2), Bio6 (Table 3), Bio8 (Table 4), Bio9 (Table 5). In addition, for 2.5m, 5m and 10m resolution rasters, Bio10 is not to be used along with Vapr5. Further, for 10m resolution raster, Bio10 is to be avoided to be used together with Vapr6 and Vapr10.

VIF pairs corresponding to Bio 11 (Mean Temperature of the Coldest Quarter)

The index of Bio 11—mean temperature of coldest quarter provides mean temperatures during the coldest three months of the year, which can be useful for examining how such environmental factors may affect species seasonal distributions. Results suggest that for four raster resolutions (30s, 2.5m, 5m and 10m), Bio11 is not to be used along with Elev, Vapr1, Vapr10, Vapr11, Vapr12, Vapr2, Vapr3, Vapr4, Vapr5, Bio1 (Table 1), Bio10 (Table 6), Bio5 (Table 2), Bio6 (Table 3), Bio8 (Table 4), Bio9 (Table 5). In addition, for 5m and 10m resolution rasters, Bio10 is not to be used along with Vapr6.

Table 3. Distribution of VIF in variable pair of Bio6 (min temperature of coldest month) with other variables corresponding to 30 arc second, 2.5 minutes, 5 minutes and 10 minutes raster resolution. VIF values with (*) indicate values <10. Reverse variable pairs of Bio6 available in other tables are not mentioned here

VarPair	30s	2.5m	5m	10m
Bio6-Bio10	26.1	29.3	33.2	39.5
Bio6-Bio11	44	48.8	52.3	58.4
Bio6-Bio8	24.7	27.4	30.2	33.9
Bio6-Bio9	31.7	33.7	35.4	38.5
Bio6-Elev	20	21.3	22.6	24.7
Bio6-Vapr1	13.3	14.9	16.6	19.3
Bio6-Vapr10	12.5	13.1	14	15.6
Bio6-Vapr11	14.7	15.7	16.8	18.7
Bio6-Vapr12	14.4	16.1	17.9	20.8
Bio6-Vapr2	13.3	14.7	16.5	18.9
Bio6-Vapr3	11.3	12.2	13.4	15
Bio6-Vapr4	13.2	14.2	15.4	17.3
Bio6-Vapr5	15.3	16.3	17.3	19.4
Bio6-Vapr6	10.3	10.9	11.6	12.8
Bio6-Vapr7	8.1*	8.6*	9.3*	10.9
Bio6-Vapr8	7.6*	8.2*	8.9*	10.4
Bio6-Vapr9	8.4*	8.8*	9.4*	10.6

Table 4. Distribution of VIF in variable pair of Bio8 (mean temperature of wettest quarter) with other variables corresponding to 30 arc second, 2.5 minutes, 5 minutes and 10 minutes raster resolution. VIF values with (*) indicate values <10. Reverse variable pairs of Bio8 available in other tables are not mentioned here

VarPair	30s	2.5m	5m	10m
Bio8-Bio10	63.8	67.8	72	80.2
Bio8-Bio11	33.6	36.5	39.7	44.4
Bio8-Bio9	20.4	22.3	24.4	28.1
Bio8-Elev	17.8	18.8	19.3	20
Bio8-Vapr1	13.5	14.9	16.3	18.8
Bio8-Vapr11	10	10.7	11.3	12.4
Bio8-Vapr12	12.1	13.2	14.3	16.3
Bio8-Vapr2	13.6	15.1	16.6	19
Bio8-Vapr3	10.4	11.3	12.3	13.8
Bio8-Vapr4	9.1*	9.9*	10.7	12.2
Bio8-Vapr5	8*	8.7*	9.5*	11

Table 5. Distribution of VIF of variable pair of Bio9 (mean temperature of driest quarter) with other variables corresponding to 30 arc second, 2.5 minutes, 5 minutes and 10 minutes raster resolution. VIF values with (*) indicate values <10. Reverse variable pairs of Bio9 available in other tables are not mentioned here.

VarPair	30s	2.5m	5m	10m
Bio9-Bio10	34.5	38.5	43.7	52.4
Bio9-Bio11	56.5	59.1	61.1	65
Bio9-Elev	19.8	20.8	21.8	23.3
Bio9-Vapr1	12.7	14.2	15.9	19.3
Bio9-Vapr10	10.4	11.1	12	13.7
Bio9-Vapr11	12	13	14	15.9
Bio9-Vapr12	13.1	14.6	16.2	19.4
Bio9-Vapr2	12.6	14.2	15.9	19.1
Bio9-Vapr3	11	12.1	13.3	15.5
Bio9-Vapr4	12.1	13.2	14.3	16.5
Bio9-Vapr5	13	13.9	14.9	17
Bio9-Vapr6	9.5*	10.2	11	12.6

VIF pairing between Bio13-Bio16 and Bio14-Bio17

Bio 13 or precipitation of wettest month identifies the total precipitation that prevails during the wettest month, which may be useful if extreme precipitation conditions during the month is known to influence potential range of a species. Bio 16 or precipitation of wettest quarter is quarterly index approximates total precipitation that prevails during the wettest three months of the year, which can be useful for examining how such environmental factors may affect species seasonal distributions.

Bio 14 or precipitation of driest month identifies the total precipitation that prevails during the driest month, which may be useful if extreme precipitation conditions during the month is known to influence potential range of a species. The quarterly index of Bio 17 or precipitation of driest quarter approximates total precipitation that prevails during the driest three months of the year, which can be useful for examining how such environmental factors may affect species seasonal distributions.

Results suggest that for four resolution (30s, 2.5m, 5m and 10m) rasters, Bio13 is not to be used along with Bio16, while Bio14 is not to be used along with Bio17 (Table 8).

VIF pairs corresponding to Elevation (Elev)

Digital Elevation Model is important raster variable used to make bioclimatic variables (Fick and Hijmans 2017). Results suggest that for four raster resolutions (30s, 2.5m, 5m and 10m), Elev is not to be used along with Vapr1, Vapr2, Vapr3, Vapr4, Vapr11, Vapr12 (Table 9), Bio1 (Table 1), Bio10 (Table 6), Bio11 (Table 3), Bio5 (Table 2), Bio6 (Table 3), Bio8 (Table 4), Bio9 (Table 5). In addition, for 2.5m, 5m, 10m resolution rasters, Elev is not to be used along with Vapr5. Further, for 10m resolution raster, Elev is not to be used along with Vapr10.

VIF pairs corresponding to Solar Radiation (Srad)

Results suggest that for four raster resolutions (30s, 2.5m, 5m and 10m), Srad10 is not to be used along with Srad11 and vice versa (Table 10).

VIF pairs corresponding to Water Vapour Pressure (Vapr)

Vapr1

For four raster resolutions (30s, 2.5m, 5m and 10m), Vapr1 or water vapour pressure for the month of January has VIF values of >10 with Vapr2, Vapr3, Vapr4, Vapr5, Vapr6, Vapr10, Vapr11, Vapr12, Bio1 (Table 1), Bio10 (Table 6), Bio11 (Table 7), Bio5 (Table 2), Bio6 (Table 3), Bio8 (Table 4), Bio9 (Table 5), and Elev (Table 9), hence Vapr1 is not to be used together with these variables. For 10m raster resolution, Vapr1 is not to be used along with Vapr7 and Vapr9.

Vapr2

For four raster resolutions (30s, 2.5m, 5m and 10m), Vapr2 or water vapour pressure for the month of February has VIF values of >10 with Vapr1, Vapr3, Vapr4, Vapr5, Vapr6, Vapr10, Vapr11, Vapr12, Bio1 (Table 1), Bio5 (Table 2), Bio6 (Table 3), Bio8 (Table 4), Bio9 (Table 5), Bio10 (Table 6), Bio11 (Table 7) and Elev (Table 9), hence Vapr2 is not to be used together with these variables. For 10m raster resolution, Vapr2 is not to be used along with Vapr7 and Vapr9.

Vapr3

For four raster resolutions (30s, 2.5m, 5m and 10m), Vapr3 or water vapour pressure for the month of March has VIF values of >10 with Vapr1, Vapr2, Vapr4, Vapr5, Vapr6, Vapr10, Vapr11, Vapr12, Bio1 (Table 1), Bio5 (Table 1), Bio6 (Table 3), Bio8 (Table 4), Bio9 (Table 5), Bio10 (Table 6), Bio11 (Table 7) and Elev (Table 9), hence Vapr3 is not to be used together with these variables. For 10m raster resolution, Vapr3 is not to be used along with Vapr7 and Vapr9.

Table 6. Distribution of VIF in variable pair of Bio10 (mean temperature of warmest quarter) with other variables corresponding to 30 arc second, 2.5 minutes, 5 minutes and 10 minutes raster resolution. VIF values with (*) indicate values <10. Reverse variable pairs of Bio10 available in other tables are not mentioned here

VarPair	30s	2.5m	5m	10m
Bio10-Bio11	37.1	40.7	45.5	53.9
Bio10-Elev	29.9	32.7	34.9	37.2
Bio10-Vapr1	17.1	19	21.1	24.9
Bio10-Vapr2	18.2	20.4	22.7	26.8
Bio10-Vapr3	13.7	15	16.4	18.9
Bio10-Vapr4	11.6	12.7	13.9	16.3
Bio10-Vapr11	12.1	12.8	13.7	15.4
Bio10-Vapr12	15.5	17	18.7	21.8
Bio10-Vapr5	9.2*	10.1	11.1	13.2
Bio10-Vapr6	6.9*	7.6*	8.4*	10.1
Bio10-Vapr10	8.3*	8.9*	9.8*	11.3

Table 7. Distribution of VIF in variable pair of Bio11 (mean temperature of coldest quarter) with other variables corresponding to 30 arc second, 2.5 minutes, 5 minutes and 10 minutes raster resolution. VIF values with (*) indicate values <10. Reverse variable pairs of Bio11 available in other tables are not mentioned here

VarPair	30s	2.5m	5m	10m
Bio11-Elev	20.2	21.1	21.7	22.6
Bio11-Vapr1	13.4	15.4	17.5	21.9
Bio11-Vapr10	10.2	11.1	12.1	14.2
Bio11-Vapr11	12.7	14	15.5	18.5
Bio11-Vapr12	13.6	15.6	17.7	22
Bio11-Vapr2	13.3	15.2	17.3	21.4
Bio11-Vapr3	11.1	12.4	13.8	16.4
Bio11-Vapr4	12.3	13.8	15.5	18.7
Bio11-Vapr5	13	14.5	16.1	19.5
Bio11-Vapr6	8.7*	9.5*	10.3	11.9

Table 8. Distribution of VIF in variable pair of Bio13-Bio16 and Bio14-Bio17 corresponding to 30 arc second, 2.5 minutes, 5 minutes and 10 minutes raster resolution

VarPair	30s	2.5m	5m	10m
Bio13-Bio16	21	21.3	21.6	22
Bio14-Bio17	78.7	77.9	77.4	78.5

Vapr4

For four raster resolutions (30s, 2.5m, 5m and 10m), Vapr4 or water vapour pressure for the month of April has VIF values of >10 with Vapr1, Vapr2, Vapr3, Vapr5, Vapr6, Vapr7, Vapr9, Vapr10, Vapr11, Vapr12, Bio1 (Table 1), Bio6 (Table 3), Bio9 (Table 5), Bio10 (Table 6), Bio11 (Table 7) and Elev (Table 9), hence Vapr4 is not to be used together with these variables. For 5m and 10m raster resolutions, Vapr4 is not to be used along with Vapr8, Bio5 (Table 2) and Bio8 (Table 4).

Vapr5

For four raster resolutions (30s, 2.5m, 5m and 10m), Vapr5 or water vapour pressure for the month of May has VIF values of >10 with Vapr1, Vapr2, Vapr3, Vapr4, Vapr6, Vapr7, Vapr8, Vapr9, Vapr10, Vapr11, Vapr12, Bio1 (Table 1), Bio11 (Table 7), Bio6 (Table 3) and Bio9 (Table 5), hence Vapr5 is not to be used together with these variables. For 2.5m, 5m and 10m raster resolutions, Vapr5 is not to be used along with Bio10 (Table 6) and Elev (Table 9). For 10m raster resolutions, Vapr5 is not to be used along with Bio5 (Table 2) and Bio8 (Table 4).

Vapr6

For four raster resolutions (30s, 2.5m, 5m and 10m), Vapr6 or water vapour pressure for the month of June has VIF values of >10 with Vapr1, Vapr2, Vapr3, Vapr4, Vapr5, Vapr7, Vapr8, Vapr9, Vapr10, Vapr11, Vapr12 and Bio6 (Table 3), hence Vapr6 is not to be used together with these variables. For 2.5m, 5m and 10m raster resolutions, Vapr6 is not to be used along with Bio9. For 5m and 10m raster resolutions, Vapr6 is not to be used along with Bio11. For 10m raster resolution, Vapr6 is not to be used along with Bio1 and Bio10.

Vapr7

For four raster resolutions (30s, 2.5m, 5m and 10m), Vapr7 or water vapour pressure for the month of July has VIF values of >10 with Vapr4, Vapr5, Vapr6, Vapr8, Vapr9, Vapr10 and Vapr11, hence Vapr7 is not to be used together with these variables. For 5m and 10m raster resolutions, Vapr7 is not to be used along with Vapr12. For 10m raster resolution, Vapr7 is not to be used along with Bio6, Vapr1, Vapr2, Vapr3.

Vapr8

For four raster resolutions (30s, 2.5m, 5m and 10m), Vapr8 or water vapour pressure for the month of August has VIF values of >10 with Vapr5, Vapr6, Vapr7, Vapr9, Vapr10 and Vapr11, hence Vapr8 is not to be used together with these variables. For 5m and 10m raster resolutions, Vapr8 is not to be used along with Vapr4. For 10m raster resolution, Vapr8 is not to be used along with Vapr12 and Bio6 (Table 3).

Vapr9

For four raster resolutions (30s, 2.5m, 5m and 10m), Vapr9 or water vapour pressure for the month of September has VIF values of >=10 with Vapr4, Vapr5, Vapr6, Vapr7, Vapr8, Vapr10 and Vapr11, hence Vapr9 is not to be used together with these variables. For 2.5m, 5m and 10m raster resolutions, Vapr9 is not to be used along with Vapr12. For 10m raster resolution, Vapr9 is not to be used along with Bio6 (Table 3), Vapr1, Vapr2 and Vapr3.

Vapr10

For four raster resolutions (30s, 2.5m, 5m and 10m), Vapr10 or water vapour pressure for the month of October has VIF values of >=10 with Vapr1, Vapr2, Vapr3, Vapr4, Vapr5, Vapr6, Vapr7, Vapr8, Vapr9, Vapr11, Vapr12, Bio11 (Table 7), Bio6 (Table 3) and Bio9 (Table 5), hence Vapr10 is not to be used together with these variables. For 2.5m, 5m and 10m raster resolutions, Vapr10 is not to be used along with Bio1 (Table 1). For 10m raster resolution, Vapr10 is not to be used along with Elev (Table 9) and Bio10 (Table 6).

Vapr11

For four raster resolutions (30s, 2.5m, 5m and 10m), Vapr11 or water vapour pressure for the month of November has VIF values of >=10 with Vapr12, Bio1 (Table 1), Bio10 (Table 6), Bio11 (Table 7), Bio6 (Table 3), Bio8 (Table 4), Bio9 (Table 5), Elev (Table 9), Vapr1, Vapr10, Vapr2, Vapr3, Vapr4, Vapr5, Vapr6, Vapr7, Vapr8 and Vapr9, hence Vapr11 is not to be used together with these variables. For 5m and 10m raster resolutions, Vapr11 is not to be used along with Bio5 (Table 2).

Vapr12

For four raster resolutions (30s, 2.5m, 5m and 10m), Vapr12 or water vapour pressure for the month of December has VIF values of >=10 with Bio1 (Table 1), Bio10 (Table 6), Bio11 (Table 7), Bio5 (Table 2), Bio6 (Table 3), Bio8 (Table 4), Bio9 (Table 5), Elev (Table 9), Vapr1, Vapr10, Vapr11, Vapr2, Vapr3, Vapr4, Vapr5 and Vapr6, hence Vapr12 is not to be used together with these variables. For 2.5m, 5m and 10m raster resolutions, Vapr12 is not to be used along with Vapr9. For 5m and 10m raster resolutions, Vapr12 is not to be used along with Vapr7. For 10m raster resolutions, Vapr12 is not to be used along with Vapr8.

Table 9. Distribution of VIF in variable pair of Elevation with other variables corresponding to 30 arc second, 2.5 minutes, 5 minutes and 10 minutes raster resolution. VIF values with (*) indicate values <10. Reverse variable pairs of Elev available in other tables are not mentioned here

VarPair	30s	2.5m	5m	10m
Elev-Vapr1	14.6	15.6	16.5	18
Elev-Vapr11	11.4	11.8	12.3	13.1
Elev-Vapr12	14.2	15.2	16.1	17.7
Elev-Vapr2	16.6	17.8	18.8	20.3
Elev-Vapr3	13.7	14.5	15.4	16.7
Elev-Vapr4	12.5	13.1	13.9	15.2
Elev-Vapr5	9.5*	10	10.5	11.7
Elev-Vapr10	8.4*	8.7*	9.1*	10

Table 10. Distribution of VIF in variable pair of Srad10 and Srad11 corresponding to 30 arc second, 2.5 minutes, 5 minutes and 10 minutes raster resolution

VarPair	30s	2.5m	5m	10m
Srad10-Srad11	13.5	13	12.7	12.7

Table 11. Distribution of VIF among variable pair of water vapour pressure variables corresponding to 30 arc second, 2.5 minutes, 5 minutes and 10 minutes raster resolution. VIF values with (*) indicate values <10. Reverse variable pairs of water vapour pressure available in other tables are not mentioned here.

VarPair	30s	2.5m	5m	10m
Vapr1-Vapr10	16.9	18.7	20.8	24.2
Vapr1-Vapr11	38.6	40.5	42.6	45.7
Vapr1-Vapr12	195.4	216.5	233.6	264.7
Vapr1-Vapr2	186.2	201.5	221.8	250
Vapr1-Vapr3	102	111.1	121.5	138.5
Vapr1-Vapr4	40.8	45.7	50.8	61.5
Vapr1-Vapr5	16.5	18.8	21.4	26.9
Vapr1-Vapr6	11.4	13.4	15.6	20.3
Vapr1-Vapr7	7*	8*	9.1*	11.6
Vapr1-Vapr9	7.8*	8.6*	9.6*	11.3
Vapr10-Vapr11	55.9	63.2	73	91.6
Vapr10-Vapr12	25.7	28.8	32.5	38.7
Vapr11-Vapr12	80.7	84.4	88.1	92.6
Vapr2-Vapr10	14.4	15.6	17.1	19.4
Vapr2-Vapr11	30.6	31.6	32.8	34.4
Vapr2-Vapr12	93	101	111.4	126.2
Vapr2-Vapr3	226.5	246.7	261.2	292.6
Vapr2-Vapr4	48.5	53.2	57.8	66.8
Vapr2-Vapr5	15.7	17.7	19.9	24.4
Vapr2-Vapr6	11.2	13	15.2	19.7
Vapr2-Vapr7	6.7*	7.5*	8.5*	10.6
Vapr2-Vapr9	7.3*	8*	8.7*	10
Vapr3-Vapr10	16.4	17.5	19	21.1
Vapr3-Vapr11	33.3	33.8	34.6	35.1
Vapr3-Vapr12	83.7	89.7	97.2	107
Vapr3-Vapr4	78.1	83.6	88.3	95.4
Vapr3-Vapr5	18.4	20.4	22.7	26.8
Vapr3-Vapr6	13.9	16.3	19.3	25.3
Vapr3-Vapr7	7.6*	8.5*	9.7*	11.9
Vapr3-Vapr9	8*	8.7*	9.4*	10.6
Vapr4-Vapr10	30.5	32.9	36.3	40.8
Vapr4-Vapr11	65.9	71.8	79	91.2
Vapr4-Vapr12	63.2	73	84.4	108.7
Vapr4-Vapr5	47.3	53.9	61.9	78
Vapr4-Vapr6	21.2	24.1	27.3	32.6
Vapr4-Vapr7	11.4	12.8	14.4	17.7
Vapr4-Vapr9	11.9	12.8	13.9	15.5
Vapr4-Vapr8	8.7*	9.6*	10.7	12.5
Vapr5-Vapr10	51.9	56	60.7	69.7
Vapr5-Vapr11	42.1	50.5	60.4	83.8
Vapr5-Vapr12	22.6	26.4	30.5	39.6
Vapr5-Vapr6	43.9	44.5	44.3	42.9
Vapr5-Vapr7	27.9	30.2	32.8	37.6
Vapr5-Vapr8	18.3	20	21.8	24.9
Vapr5-Vapr9	20.8	21.9	23.1	25.1
Vapr6-Vapr10	28.4	30.9	33	36.2
Vapr6-Vapr11	16.7	18.9	21	24.3
Vapr6-Vapr12	13.5	15.9	18.6	24
Vapr6-Vapr7	48.5	50.3	52.3	56
Vapr6-Vapr8	16.5	16.9	17.2	17.4
Vapr6-Vapr9	21.4	22.1	22.8	23.6
Vapr7-Vapr10	23.4	26.5	30.1	38.1
Vapr7-Vapr11	12	13.8	15.8	20.1
Vapr7-Vapr8	49.9	51.1	52.2	53.7
Vapr7-Vapr9	37	39.2	41.6	47
Vapr7-Vapr12	8.3*	9.6*	11	14.1
Vapr8-Vapr10	20.6	23.7	27	33.7
Vapr8-Vapr11	10.1	11.7	13.5	17.4
Vapr8-Vapr9	60	65.7	70.4	79.9
Vapr8-Vapr12	6.8*	7.7*	8.7*	10.7
Vapr9-Vapr10	45.3	50.8	57	67.3
Vapr9-Vapr11	15.4	17.3	19.5	23.5
Vapr9-Vapr12	9.7*	10.9	12.2	14.6

VIF pairs corresponding to Wind Speed (Wind)

For four raster resolutions (30s, 2.5m, 5m and 10m), Wind1 or wind speed for the month of January has VIF values of >=10 with Wind2, and conversely Wind2 (wind speed for the month of February) has VIF values of >=10 with Wind1. Similarly, for all of these data resolutions, i) Wind2 has VIF values of >=10 with Wind3 (wind speed for the month of March) and vice versa, ii) Wind3 has VIF values of >=10 with Wind4 (wind speed for the month of April) and vice versa, iii) Wind4 has VIF values of >=10 with Wind5 (wind speed for the month of May) and vice versa, iv) Wind5 has VIF values of >=10 with Wind6 (wind speed for the month of June) and vice versa, hence these variable pair combinations should be avoided (Table 11).

For the raster resolutions (5m and 10m), i) Wind2 has VIF values of >=10 with Wind12 (wind speed of December) and vice versa, and ii) Wind4 has VIF values of >=10 with Wind10 (wind speed of October) and vice versa, hence these variable pair combinations should be avoided.

For the 10m raster resolution, i) Wind3 has VIF values of >=10 with Wind5 and vice versa, and ii) Wind6 has VIF values of >=10 with Wind7 (wind speed of July) and vice versa, iii) Wind9 (wind speed of September) has VIF values of >=10 with Wind10 (wind speed of October) and vice versa, hence these variable pair combinations should be avoided.

Comparison with VIF output by *usdm* package

The VIF pairs identified for studied explanatory variables in Indonesian extent from present study was compared with the VIF of the said explanatory variables for the same extent derived from *vifstep* function of *usdm* package. The output of *vifstep* function provides total number along with names of variables from the input variables that have collinearity problem, but it doesn't provide the variable pairs having the collinearity problem. After excluding the collinear variables, the function provides value for variable pair with minimum and maximum values of linear correlation coefficients. Next, it provides VIFs of the remained variables, but here VIF is associated with a particular variable, and not the variable pair (Table 12).

The *usdm* package output enlisted 38 variables for 30s and 2.5m resolution and 39 variables for 5m and 10m resolution to have problematic collinearity (VIF>10) (Table 13). The output also mentioned i) Bio7, Bio8, Srad8 and Srad 12 for all resolutions, ii) Srad2 and Srad6 for 30s, 2.5m and 5m resolutions, and iii) Srad3 for 10m resolution to have VIF>10, but from the present study, said variables were found to have VIF<10 (Tables 13,14).

Interestingly, there were disparity between the VIF status of some 'non-collinear' variables resulting from *usdm* and their collinearity indicated from present study. Viz. for 30s resolution rasters, *usdm* ascribed VIF of 3.46 to Bio13 and VIF of 4.28 to Vapr8, indicating their non-collinearity, but for the said resolution, present study found that Bio13 is linked to Bio16 with VIF 21, and Vapr8 is linked to Vapr9 (VIF 60), Vapr10 (VIF 20.6), Vapr11 (VIF 10.1). Similarly, for 2.5m resolution rasters, *usdm* ascribed

VIF of 3.47 to Bio13 and VIF of 7.14 to Wind1, but for the said resolution, present study found that Bio13 is linked to Bio16 with VIF 21.3, and Wind1 is linked to Wind2 (VIF 10.6). For 5m resolution rasters, *usdm* ascribed VIF of 3.53 to Bio13, 4.83 to Vapr8 and 7.95 to Wind1, but the present study revealed for the said resolution, Bio13 to be linked to Bio16 with VIF 21.6, Vapr8 to be linked to Vapr9 (VIF 70.4), Vapr10 (VIF 27), Vapr11 (VIF 13.5) and Wind1 to be linked to Wind2 with VIF 11.9. In case of 10m resolution rasters, *usdm* attributed VIF of 3.47 to Bio13, 5.05 to Vapr8 and 9.96 to Wind1, but the present study revealed for the said resolution, Bio13 to be linked to Bio16 with VIF 22, Vapr8 to be linked to Vapr9 (VIF 79.9), Vapr10 (VIF 33.7), Vapr11 (VIF 17.4) and Vapr12 (VIF 10.7), while Wind1 was found to be associated with Wind2 with VIF 13.8.

Effect of data (raster) resolution

At equator (0° latitude), 30 arc-seconds (0.5-minute) spatial resolution corresponds to about 0.86 km^2 cell size (commonly referred to as '1-km' spatial resolution), 2.5 minutes raster resolution is equivalent to around 21.44 km^2 cell size, 5 minutes raster resolution is equivalent to around 85.75 km^2 cell size and 10 minutes raster resolution is equivalent to around 342.99 km^2 cell size. Species distribution are scale dependent hence it is pertinent to understand the changes in VIF values corresponding to data (raster) resolution.

Table 12. Distribution of VIF among variable pair of wind variables corresponding to 30 arc second, 2.5 minutes, 5 minutes and 10 minutes raster resolution. VIF values with (*) indicate values <10

VarPair	30s	2.5m	5m	10m
Wind1-Wind2	8.6*	10.6	11.9	13.8
Wind2-Wind3	8.9*	10.9	12.4	13.7
Wind3-Wind4	8.8*	10.8	12.3	14.5
Wind4-Wind5	8.4*	10.4	12.3	16.4
Wind5-Wind6	8.3*	10.1	11.5	13.8
Wind2-Wind12	7.7*	9.3*	10.3	11.9
Wind4-Wind10	7.3*	9.2*	10.7	12.5
Wind3-Wind5	5.9*	7.1*	8.4*	11.2
Wind6-Wind7	7.7*	8.9*	9.6*	10.7
Wind9-Wind10	6.9*	8.6*	9.6*	10.9

Table 13. Variables with collinearity problem for Indonesian extent derived from *vifstep* function of *usdm* package. Variables which were found to have VIF>10 in usdm output but having VIF<10 in the present study are bold italicized and (*) marked

30s	2.5m	5m	10m
Bio1,	Bio1,	Bio1,	Bio1,
Bio10,	Bio10,	Bio10,	Bio10,
Bio11,	Bio11,	Bio11,	Bio11,
Bio12*,	***Bio12****,	***Bio12****,	***Bio12****,
Bio14,	Bio14,	Bio14,	Bio14,
Bio16,	Bio16,	Bio16,	Bio16,
Bio17,	Bio17,	Bio17,	Bio17,
Bio5,	Bio5,	Bio5,	Bio5,
Bio6,	Bio6,	Bio6,	Bio6,
Bio7*,	***Bio7****,	***Bio7****,	***Bio7****,
Bio8,	Bio8,	Bio8,	Bio8,
Bio9,	Bio9,	Bio9,	Bio9,
Elev,	Elev,	Elev,	Elev,
Srad10,	Srad10,	Srad10,	Srad10,
Srad11,	Srad11,	Srad11,	Srad11,
Srad12*,	***Srad12****,	***Srad12****,	***Srad12****,
Srad2*,	***Srad2****,	***Srad2****,	***Srad3****,
Srad6*,	***Srad6****,	***Srad6****,	Srad7,
Srad8*,	***Srad8****,	***Srad8****,	***Srad8****,
Vapr1,	Vapr1,	Vapr1,	Vapr1,
Vapr10,	Vapr10,	Vapr10,	Vapr10,
Vapr11,	Vapr11,	Vapr11,	Vapr11,
Vapr12,	Vapr12,	Vapr12,	Vapr12,
Vapr2,	Vapr2,	Vapr2,	Vapr2,
Vapr3,	Vapr3,	Vapr3,	Vapr3,
Vapr4,	Vapr4,	Vapr4,	Vapr4,
Vapr5,	Vapr5,	Vapr5,	Vapr5,
Vapr6,	Vapr6,	Vapr6,	Vapr6,
Vapr7,	Vapr7,	Vapr7,	Vapr7,
Vapr9,	Vapr9,	Vapr9,	Vapr9,
Wind10,	Wind10,	Wind10,	Wind10,
Wind12,	Wind12,	Wind12,	Wind12,
Wind2,	Wind2,	Wind2,	Wind2,
Wind3,	Wind3,	Wind3,	Wind3,
Wind4,	Wind4,	Wind4,	Wind4,
Wind5,	Wind5,	Wind5,	Wind5,
Wind6,	Wind6,	Wind6,	Wind6,
Wind9	Wind9	Wind7,	Wind7,
		Wind9	Wind9

Table 14. The highest VIF value of the variables for Indonesian extent taken as collinear by *usdm* package but found to be non-collinear in present study

30s		2.5m		5m		10m	
VarPair	**VIF**	**VarPair**	**VIF**	**VarPair**	**VIF**	**VarPair**	**VIF**
Bio7-Bio2	1.8	**Bio7**-Bio2	1.8	**Bio7**-Bio2	1.9	**Bio7** – Vapr8	1.3
Bio12-Bio17	4.9	**Bio12**-Bio17	4.7	**Bio12**-Bio17	4.6	**Bio12**-Bio17	4.6
Srad8-Srad7	3.3	**Srad8**-Srad7	3.3	**Srad8**-Srad7	3.4	**Srad8**-Srad7	3.6
Srad6-Srad7	3.1	**Srad6**-Srad7	2.9	**Srad6**-Srad7	2.8		
Srad2-Srad3	7.7	**Srad2**-Srad3	8.1	**Srad2**-Srad3	8.6	**Srad3**-Srad2	9
Srad12-Srad11	8.5	**Srad12**-Srad11	7.9	**Srad12**-Srad11	7.5	**Srad12**-Srad11	6.9

It was observed from the summary statistics (Table 15) that going from the gradient of fine resolution (30s) towards coarser data resolution (10 m), there was increase in variance, increase in large value VIFs (outliers), increase in median, mean and corresponding standard deviation of mean. Figure 1 depicts visual comparison of four raster cell sizes with example of mean temperature of warmest quarter (Bio10) map in Indonesia, with zoomed inset of Gunung Leuser National Park region, Northern Sumatra, which will be helpful to understand the grain size/ raster resolution at the scale of Indonesia and in preliminarily selection of preferred raster resolution as per study objective.

To study the significance of pair-wise raster differences, firstly, data distribution (VIF) was assessed with Shapiro-Wilk normality test which resulted in W = 0.54311, p-value < 2.2e-16, meaning that p-value <0.05 showed strong evidence of the data being non-normal. It was further corroborated with the density distribution plot of raster resolution-wise VIF, which showed strong skewed values towards left side (Figure 2), with the presence of large value VIFs as outliers (Figure 3).

As the distribution was found to be non-normal, Kruskal-Wallis rank sum test (non-parametric version of One-way ANOVA) and Friedman rank sum test (non-parametric version of one-way repeated measures ANOVA) were carried out.

Kruskal-Wallis rank sum test showed chi-squared = 29.203, df = 3, p-value = 2.03e-06 and Friedman rank sum test showed chi-squared = 494.47, df = 3, p-value < 2.2e-16. As both the results showed p-value <0.05, the null hypothesis was rejected and significant difference between VIF values of raster resolutions were considered. To understand particularly the pairs of different individual raster resolutions, post hoc Wilcoxon rank sum test with pair-wise comparisons with bonferroni continuity correction was conducted. It was observed that 30s resolution raster was significantly different (p-value <0.05) than 5m and 10m, while 2.5m resolution raster was significantly different (p-value <0.05) than 10m resolution raster (Table 16).

Table 15. Summary statistics of resolution-wise variance inflation factor values

Res	Min.	Max.	Mean	SD	Median	Variance
30s	5.90	226.50	26.68	33.81	13.80	1142.95
2.5m	7.10	246.70	29.16	36.90	15.60	1361.27
5m	8.40	271.60	31.87	40.29	17.15	1623.43
10m	10.00	330.90	36.68	46.36	19.85	2149.37

Table 16. Post hoc Wilcoxon rank sum test with pair-wise comparisons of rasters with *bonferroni* continuity correction

Resolution	30s	2.5m	5m
2.5m	0.6588	-	-
5m	0.0155	0.7003	-
10m	4.50E-06	0.0011	0.0917

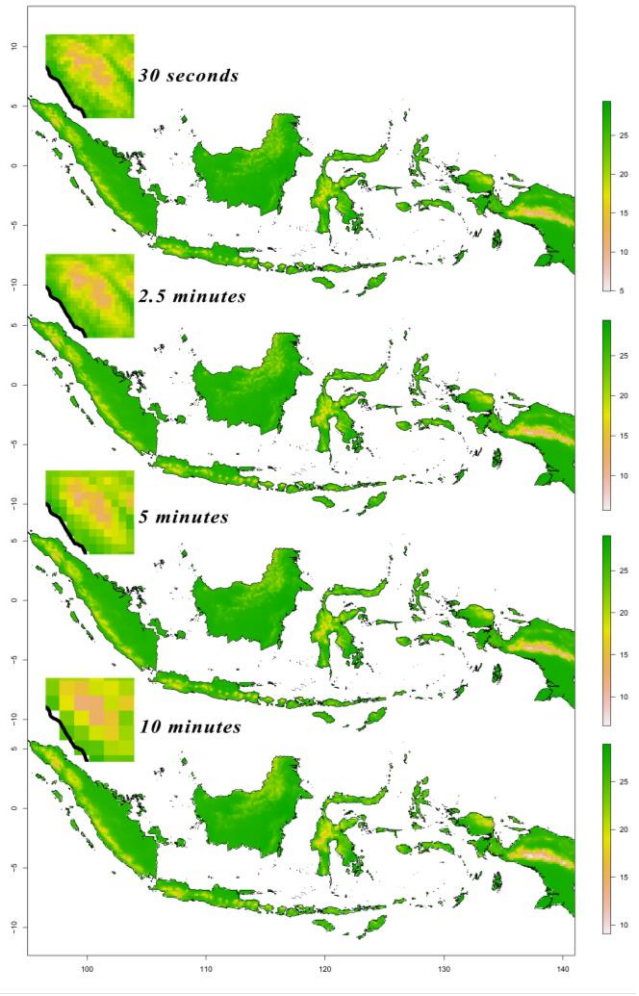

Figure 1. Visual comparison of four raster cell sizes with example of mean temperature of warmest quarter (Bio10) map in Indonesia, with zoomed inset of Gunung Leuser National Park region, northern Sumatra.

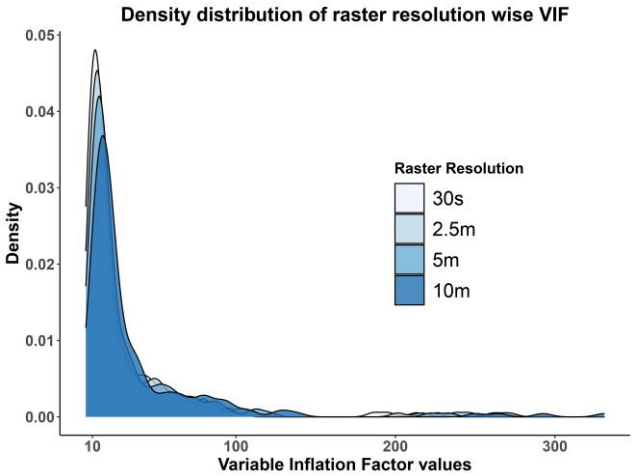

Figure 2. Density distribution of VIF with respect to various raster resolution

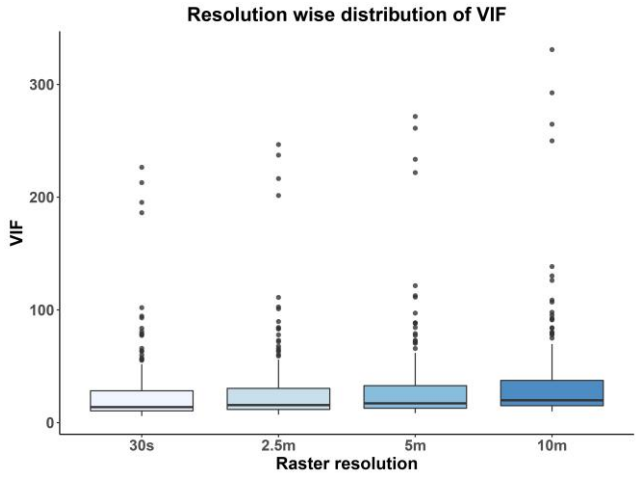

Figure 3. Distribution of VIF with respect to various raster resolution

Tentative list of non-collinear variables

For all the raster resolutions, Bio2, Bio3, Bio4, Bio15, Bio18, Bio19, slope, aspect, Srad1, Srad4, Srad5, Srad9, Wind8, Wind11 were found to be non-collinear. While, Srad3 and Srad 7 were found to be non-collinear for 30s, 2.5m and 5m raster resolutions; Wind 7 was non-collinear for 30s and 2.5m; Srad2 and Srad6 were non-collinear for 10m; Vapr8 for 2.5m and Wind1 was non-collinear for 30s raster resolutions.

Discussion

From the analysis of 58 explanatory/predictor variables and their corresponding 3364 variable pairs, 174 resolution wise variable pairs were known to be affected by multicollinearity, out of which temperature related bioclimatic variables, water vapour pressure and elevation associated variables were highly notable. Resolution-wise, 10m rasters had the highest variance, indicating more data noise and presence of greater number of collinear variables, while 30s rasters had the lowest variance, meaning these datasets deviate less significantly than mean, and have a smaller number of collinear variables.

Regarding bioclimatic variables, temperature variables mainly affected by multicollinearity include mean temperatures of warmest (Bio10), wettest (Bio8), coldest (Bio11) and driest quarter (Bio9), annual mean temperature (Bio1), as well as limiting factors like maximum temperature of warmest month (Bio5) and minimum temperature of coldest month (Bio6). With respect to the precipitation related factors, precipitation of wettest month (Bio13) was found to be collinear with precipitation of wettest quarter (Bio16). Further, precipitation of driest month (Bio14) was found to be collinear with precipitation of driest quarter (Bio17). Water vapour pressure also had high multicollinearity among themselves, with elevation and temperature related bioclimatic variables as water vapour pressure itself is calculated from dew-point temperature/ mean relative humidity and mean temperature (Fick and Hijmans 2017). Higher multicollinearity of elevation with temperature related variables may be due to dependence of temperature with gradients of latitude and elevation, and less multicollinearity among precipitation related variables may be explained by the fact that precipitation can be highly variable in time and space and some regions have abrupt changes (Fick and Hijmans 2017). Wind speed was found to be interesting in a sense that none of the 30 second rasters were multicollinear (VIF values <10), while gradually VIF values of concerned variable pairs increased with the increase in resolution.

While selecting explanatory variables for ecological niche modeling, composite variables based on the precipitation of the coldest or warmest period or temperature of the driest or wettest period could be avoided as these datasets are hinted to be internally flawed. Therefore, limiting factors like the maximum temperature of the warmest month, minimum temperature of the coldest period, temperature variability, precipitation variability, precipitation of the wettest and driest periods and so on may be used in combination unless otherwise VIF values restrict them (Pradhan 2016). However, working only with non-redundant ones may not always yield good results as some of the redundant variables may act as good ecological descriptors, which could be used in ENM process without its corresponding variable pair (VIF>10). It should also be noted that sometimes overfitting may be the result of selecting too much aggregate sampling sites (sample/observation bias) which may be corrected (Pradhan 2016).

After VIF screening, the variables may be preliminarily run in MaxEnt (minimum of triplicate runs), which may be helpful to identify the least significant variables based on jackknife test for evaluating relative importance of variables, % contribution to the model as well as individual response to the variable (Jueterbock et al. 2016; Gunawan et al. 2021). In case, multiple models are built for the same species utilizing multiple sets of non-collinear variables, the final model is suggested to selected based upon lowest AICc score, highest AUC value and incorporating lesser number of correlated variables (VIF <10) (Warren et al. 2010). *ENMeval* (Muscarella et al. 2014; Kass et al. 2021) and *Maxentvariableselection* packages in R offer various evaluation metrics for selecting explanatory variables which may be consulted prior to model building (Jueterbock et al. 2016) and the overall workflow of ecological niche modeling may be conducted following Overview, Data, Model, Assessment and Prediction (ODMAP) protocol outlined by Zurell et al. (2020).

Further, under current climatic conditions, performance and prediction likelihood of models based on CHELSA climatic database are reported to outperform than that of WorldClim climatic database, especially for high mountain regions (Bobrowski 2021). Besides, it may also be kept in mind that all temperature variables of WorldClim 2.1 are based upon the higher global correlation coefficient (ρ) between estimated and observed values of 0.99, and similarly solar radiation and water vapour pressure both have correlation coefficients higher than 0.95; however, accuracy was lowest for wind speed (ρ=0.76) and precipitation (ρ=0.86) (Fick and Hijmans 2017), which is

important in understanding how realistic explanatory variables are we using for ENM studies as regional/ local scale (Pradhan 2019).

In conclusion the study presented a primer for selection of various explanatory/predictor variables based upon WorldClim 2.1 datasets available in four raster resolutions. At the extent of Indonesia, out of 58 explanatory variables and their corresponding 3364 variable pairs, 174 variable pairs were known to be affected by multicollinearity, from which temperature related bioclimatic variables, water vapour pressure and elevation associated variables were highly notable. For all the raster resolutions, Bio2, Bio3, Bio4, Bio15, Bio18, Bio19, slope, aspect, Srad1, Srad4, Srad5, Srad9, Wind8, Wind11 were found to be non-collinear. While, Srad3 and Srad 7 were found to be non-collinear for 30s, 2.5m and 5m raster resolutions; Wind 7 was non-collinear for 30s and 2.5m; Srad2 and Srad6 were non-collinear for 10m; Vapr8 for 2.5m and Wind1 was non-collinear for 30s raster resolutions. In a gradient smaller, resolution raster had smaller data variance e.g. 30s than the larger resolution raster e.g. 10m. VIF output of *usdm* package of R was compared with present study, and some disparities were noted, necessitating validation of VIF of screened variables who rely solely on such packages. Besides WorldClim, other climatic databases such as CHELSA is to be compared and explored for regional ENM studies.

ACKNOWLEDGEMENTS

Authors acknowledge the precise perusal of the manuscript and thoughtful comments by the anonymous reviewers which were very helpful in improvement of the manuscript. First author is thankful to the Indian Statistical Institute, Kolkata for their 2018 workshop on economic and ecological impacts of Invasive Alien Species, during the course of which possibilities of R program was introduced; further, the first author is indebted to miscellaneous insights in R and R codes available at Stackoverflow threads (https://stackoverflow.com/), GIS Stackexchange threads (https://gis.stackexchange.com/), Statistical tools for high-throughput data analysis (http://www.sthda.com/english/) and Facebook group *Ecology in R*.

REFERENCES

Anderson RP, Gonzales I. 2011. Species-specific tuning increases robustness to sampling bias in models of species distributions: An implementation with Maxent. Ecol Modell 222: 2796–2811. DOI: 10.1016/j.ecolmodel.2011.04.011.

Bivand R, Keitt T, Rowlingson B. 2020. rgdal: Bindings for the 'Geospatial' Data Abstraction Library. R package version 1.5-16. https://CRAN.R-project.org/package=rgdal

Bobrowski M, Weidinger J, Schickhoff U. 2021. Is New Always Better? Frontiers in Global Climate Datasets for Modeling Treeline Species in the Himalayas. Atmosphere 12: 543. DOI: 10.3390/atmos12050543.

Chang W. 2020. webshot2: Takes screenshots of web pages, including Shiny applications and R Markdown documents. https://github.com/rstudio/webshot2.git

EUROSTAT 2020. Global country boundaries for 2020. Available at https://gisco-services.ec.europa.eu/distribution/v2/countries/download/ref-countries-2020-01m.shp.zip, accessed on 22.05.2021.

Faraway J. 2016. faraway: Functions and Datasets for Books by Julian Faraway. R package version 1.0.7. https://CRAN.R-project.org/package=faraway.

Fick SE, Hijmans RJ. 2017. Worldclim 2: New 1-km spatial resolution climate surfaces for global land areas. International Journal of Climatology. DOI: 10.1002/joc.5086.

Fox J, Weisberg S. 2019. An {R} Companion to Applied Regression, Third Edition. Sage Publishing . Los Angeles, USA.

Galili T, O'Callaghan A, Sidi J, Sievert C. 2017. heatmaply: an R package for creating interactive cluster heatmaps for online publishing, Bioinformatics, btx657, DOI: 10.1093/bioinformatics/btx657.

Gunawan, Rizki MI, Anafarida O, Mahmudah N. 2021. Modeling potential distribution of Baccaurea macrocarpa in South Kalimantan, Indonesia. Biodiversitas 22: 3230-3236. DOI: 10.13057/biodiv/d220816

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A. 2005. Very high resolution interpolated climate surfaces for global land areas. Int J Climatol 25: 1965-1978. DOI: 10.1002/joc.1276.

Hijmans RJ. 2020. raster: Geographic Data Analysis and Modeling. R package version 3.4-5. https://CRAN.R-project.org/package=raster

Jueterbock A, Smolina I, Coyer JA, Hoarau G. 2016. The fate of the Arctic seaweed Fucus distichus under climate change: an ecological niche modeling approach. Ecol Evol 6n(6): 1712–1724. DOI: 10.1002/ece3.2001 R.

Kass JM, Muscarella R, Galante PJ, Bohl CL, Pinilla-Buitrago GE, Boria RA, Soley-Guardia M, Anderson RP. 2021. ENMeval 2.0: Redesigned for customizable and reproducible modeling of species' niches and distributions. Methods Ecol Evol 00: 1-7. DOI: 10.1111/2041-210X.13628.

Muscarella R, Galante PJ, Soley-Guardia M, Boria RA, Kass J, Uriarte M, Anderson RP. 2014. ENMeval: An R package for conducting spatially independent evaluations and estimating optimal model complexity for ecological niche models. Methods Ecol Evol 5 (11): 1198-1205. DOI: 10.1111/2041-210X.12261.

Naimi B, Hamm Na, Groen TA, Skidmore AK, Toxopeus AG. 2014. Where is positional uncertainty a problem for species distribution modelling. Ecography 37: 191-203. DOI: 10.1111/j.1600-0587.2013.00205.x.

Nursamsi I, Partasasmita R, Cundaningsih N, Ramadhani HS. 2018. Modeling the predicted suitable habitat distribution of Javan hawk-eagle Nisaetus bartelsi in the Java Island, Indonesia. Biodiversitas 19: 1539-1551. DOI: 10.13057/biodiv/d190447.

O'Donnell MS, Ignizio DA. 2012. Bioclimatic predictors for supporting ecological applications in the conterminous United States: U.S. Geological Survey Data Series 691.

Pradhan P. 2015. Potential distribution of *Monotropa uniflora* L. as a surrogate for range of Monotropoideae (Ericaceae) in South Asia. Biodiversitas 16 (2): 109-115. DOI: 10.13057/biodiv/d160201

Pradhan P. 2016. Strengthening MaxEnt modelling through screening of redundant explanatory bioclimatic variables with variance inflation factor analysis. Researcher 8(5): 29-34. DOI: 10.7537/marsrsj080516.05

Pradhan P. 2019. Testing equivalency of interpolation derived bioclimatic variables with actual precipitation: A step towards selecting more realistic explanatory variables for species distribution modelling. Res J Chem Environ 23: 38-41.

Setyawan AD, Supriatna J, Darnaedi D, Rokhmatuloh, Sutarno, Sugiyarto, Nursamsi I, Komala W, Pradan P. 2017. Impact of climate change on potential distribution of xero-epiphytic Selaginellas (Selaginella involvens and S. repanda) in Southeast Asia. Biodiversitas 18 (4): 1680-1695. DOI: 10.13057/biodiv/d180448.

Setyawan AD, Supriatna J, Nisyawati, Nursamsi I, Sutarno, Sugiyarto, Sunarto, Prdan P, Budiharta S, Pitoyo A, Suhardono S, Setyono P, Indrawan M. 2020a. Predicting potential impacts of climate change on the geographical distribution of mountainous selaginellas in Java, Indonesia. Biodiversitas 21 (10): 4866-4877. DOI: 10.13057/biodiv/d211053.

Setyawan AD, Supriatna J, Nisyawati, Nursamsi I, Sutarno, Sugiyarto, et al. 2020b. Anticipated climate changes reveal shifting in habitat suitability of high-altitude selaginellas in Java, Indonesia. Biodiversitas 21 (11): 5482-5497. DOI: 10.13057/biodiv/d211157.

Setyawan AD, Supriatna J, Nisyawati, Nursamsi I, Sutarno, Sugiyarto, et al. 2021. Projecting expansion range of Selaginella zollingeriana in

the Indonesian archipelago under future climate conditions. Biodiversitas 22 (4): 2088-2103. DOI: 10.13057/biodiv/d220458.

Suwarto, Prasetyo LB, Kartono AP. 2016. Habitat suitability for Proboscis Monkey (Nasalis larvatus Wurmb, 1781) in the mangrove forest of Kutai National Park, East Kalimantan. Bonorowo Wetlands 6: 12-25. DOI: 10.13057/bonorowo/w060102.

Vaidyanathan R, Xie Y, Allaire JJ, Cheng J, Russell K. 2019. htmlwidgets: HTML Widgets for R. R package version 1.5.1. https://CRAN.R-project.org/package=htmlwidgets.

Warren DL, Glor RE, Turelli M. 2010. ENM Tools: a toolbox for comparative studies of environmental niche models. Ecography 33: 607–611. DOI: 10.1111/j.1600-0587.2009.06142.x.

Yudaputra A, Pujiastuti I, Cropper Jr. WP. 2019. Comparing six different species distribution models with several subsets of environmental variables: predicting the potential current distribution of Guettarda speciosa in Indonesia. Biodiversitas 20: 2321-2328. DOI: 10.13057/biodiv/d20083.

Zurell D, Franklin J, König C, Bouchet PJ, Dormann CF, Elith J, Fandos G, Feng X, Guillera-Arroita G, Guisan A, Lahoz-Monfort JJLeitao PJ, Park DS, Peterson AT, Rapacciuolo G, Schmatz DR, Schroder B, Serra-Diaz JM, Thuiller W, Yates KL, Zimmermann NE, Merow C. 2020. A standard protocol for reporting species distribution models. Ecography 43: 1261-1277. DOI: 10.1111/ecog.04960.