New 30 m resolution Hong Kong climate, vegetation, and topography rasters indicate greater spatial variation than global grids within an urban mosaic

Brett Morgan[1] and Benoit Guénard[1]

[1]School of Biological Sciences, The University of Hong Kong, Hong Kong SAR
Correspondence: Brett Morgan (brettmorgan2@gmail.com)

Abstract. The recent proliferation of high quality global gridded GIS datasets has spurred a renaissance of studies in many fields, particularly biogeography. However these data, often 1 km at the finest scale available, are too coarse for applications such as precise designation of conservation priority areas and species distribution modeling, or purposes outside of biology such as city planning and precision agriculture. Further, these global datasets likely underestimate local climate variations because they do not incorporate locally relevant variables. Here we describe a comprehensive set of 30 m resolution rasters for Hong Kong, a small subtropical territory with highly variable terrain where intense anthropogenic disturbance meets a robust protected area system. The data include topographic variables, Normalized Difference Vegetation Index, and interpolated climate variables based on weather station observations. We present validation statistics that convey each climate variable's reliability, and compare our results to a widely used global dataset, finding that our models consistently reflect greater climatic variation. To our knowledge, this is the first set of published environmental rasters specific to Hong Kong. We hope this diverse suite of geographic data will facilitate future environmental and ecological studies in this region of the world, where a spatial understanding of rapid urbanization, introduced species pressure, and conservation efforts is critical. The dataset is accessible at https://figshare.com/s/3a5634e36e80dc33444c.

1 Introduction

Scale of analysis has long been considered a key concern in biogeographic research (Levin, 1992). Multiple types of scale are relevant to environmental data, including analysis grain, response grain, spatial structure, and study extent (Mertes and Jetz, 2018). Analysis grain, the minimum unit of spatial resolution in a spatial grid, is commonly referred to as a pixel or cell. In research that uses environmental raster data, the pixel size directly dictates the types of biogeographic questions that can be reasonably addressed.

This relationship between analysis grain and study suitability is complex, and higher resolutions are not always advantageous. For example, in global analyses excessively high resolution data would be computationally cumbersome and unnecessary if the goal is to characterize broad patterns. However as shown below, many studies have found notable benefits of higher resolution climatic predictors. Unfortunately, regional analyses lacking local data are limited to

using global datasets and the grain size at which they are available (e.g. Cheng and Bonebrake, 2017).

Species distribution modeling (SDM) is a common application of gridded environmental data, where the selected analysis grain has important consequences. In SDM, one or more geographic predictors are associated statistically with the location of known observations of a species. The resulting statistical model can be converted to a geographic model: a spatially continuous measure of species occurrence likelihood across the landscape of interest. SDMs are used for many applications, including predicting potential ranges of invasive species, characterizing ecological constraints on species ranges, discovering biodiversity, and planning protected areas (Peterson et al., 2011). The effects of SDM grain size manipulation is an active area of research. Below, we summarize findings on four main effects; estimated distribution size, inclusion of fine scale features, predictor variable selection, and model predictive ability.

Coarser environmental data consistently result in SDMs that predict larger areas of species presence (Connor et al., 2017; Franklin et al., 2013; Seo et al., 2009). Overestimation of SDMs is especially a concern for conservation purposes, where inferred size of suitable habitat is often used to inform extinction risk assessments. Mistakenly large calculated distributions could result in species that are assigned artificially low risk levels.

Coarse resolution predictors can cause SDMs to omit small, but important areas. Particularly of interest are microrefugia, climatically unique patches of land that can harbor rare species, and are especially important for conservation as species distributions respond to climate change (Dobrowski, 2010). Meineri and Hylander (2017) demonstrated that because high resolution climate models included such microrefugia, the resulting species distribution models predicted lower extinction rates for plant species than coarser predictors. Nezer et al. (2016) found that 10 m or 100 m resolution SDMs can reveal other distribution features invisible at lower resolutions (1 km): movement corridors, isolated habitat patches, geomorphologic features, and anthropogenic effects on distributions.

SDM scale can also affect which predictors are selected for model calculation. Certain predictors may be excluded in SDMs because they lack explanatory power at the chosen scale of analysis (Mertes and Jetz, 2017). For example, vegetation measures like the Normalized Difference Vegetation Index (NDVI) in fragmented forests are unlikely to be relevant if the grain size is much larger than the forest patch size, because each grid cell will be a single averaged value. This means that coarse models might not only mischaracterize the distribution pattern itself, but they also may fail to explicate important environmental relationships that determine species occurrence. Indeed, Nezer et al. (2016) found that the most important predictors (vegetation, slope) in their highest resolution models (10 m) were "nearly meaningless" at 1 km resolution. Another study found similar differences in predictor importance related to variation in scale (Lasseur et al., 2006). Of course, predictor importance is always relative and thus is subject to which predictors are included in model building. Therefore this pattern is not expected to be observed in all studies, but should not be overlooked as a potential source of bias.

Last, any consistent effects of SDM grain size on the overall predictive ability of SDMs are unclear. The most commonly used measure of SDM performance is Area Under Curve (AUC), where a higher value indicates a greater ability to differentiate between area the species is present or absent. Some studies found increased SDM resolution resulted in increased AUC (Seo et al., 2009; Nezer et al., 2016), while others found no effect (Pradervand et al. 2014) or mixed effects depending on dataset (Guisan et al., 2007). These studies used different species, predictors, scales, regions, and modeling algorithms, so further research is required to investigate any association between SDM grain size and AUC.

The above advantages of higher resolution environmental data in SDM may be dependent on project-specific factors, such as the quality of species records available and the goals of the research. For example, using environmental grids of a smaller grain size than the locational accuracy of the available species records is untenable. Additionally, stationary species (e.g. lichens) may be more strongly affected by local factors while highly mobile species (e.g. birds) may only be limited at broader scales. Indeed, it has been shown that plant (rather than bird or mammal) species models with highest locational accuracy were those most improved by higher resolution (Guisan et al., 2007). Lastly, the utility of fine grain environmental grids may depend on habitat; flat deserts likely have less biologically relevant fine-scale spatial variation compared to mountainous forests or subtropical areas fragmented by human activity, like Hong Kong.

In this study, a new series of rasters for Hong Kong are introduced particularly suited for SDM. The layers produced focus on long term climate averages, topography, and vegetation. We asked how the new 30 m scale rasters provide new information on climatic variables in Hong Kong, in comparison to a global dataset already available. We hypothesize that our new climate data will indicate greater variation (measured as raster standard deviation) in climate variables. The development of high-resolution environmental rasters is particularly important in tropical regions where species exhibit small distribution ranges (as predicted by Rapoport's Rule: Stevens, 1989) and where understanding interactions between organisms and their changing habitats is paramount.

2 Study area: Hong Kong

Geographic data of appropriate resolution is critically important for conducting research within the Hong Kong Special Administrative Region of China, because of its complex landscape. Hong Kong exhibits dramatically variable topography, fitting numerous small islands, dozens of mountain peaks over 500 m, 733 km of coastline, and a human population of over 7 million into a land area of only 1,104 km2 (Fig. 1). Seasonally variable monsoon winds deliver equatorial heat and torrential precipitation in summer, while northerly winds carry chilly dry air from continental Asia during the winter (Dudgeon and Corlett, 1994). However, daily temperature fluctuations are attenuated by the surrounding South China Sea and Pearl River Estuary. Hong Kong's terrain typically exhibits a stark bifurcation between some of the most densely constructed areas in the world (Lau and Zhang, 2015) and steep, vegetated slopes. Uninhabited expanses are protected as part of 24 country parks and additional special areas that cover over 40% of the territory's land (Agriculture, Fisheries and Conservation Department, 2017). Even

within these more natural areas, a strong disturbance gradient encompasses grasslands, shrublands, evergreen secondary forests, and old-growth feng shui woods that have been protected from deforestation. Historically Hong Kong has been largely stripped of its trees, and only since the end of World War II and later the establishment of the Country Park system have large swathes of forest begun to regenerate (Zhuang and Corlett, 1997). However this process is frequently reset by human-induced hill fires, which maintain predominantly upland areas as shrubland or grassland (Marafa and Chau, 1999). Hong Kong harbors several unique and restricted habitats, including mangroves in coastal areas and freshwater wetlands in the far northwest.

Hong Kong climate data is available within a variety of global gridded climate datasets (WorldClim 2 - Fick and Hijmans, 2017; MerraClim - Vega et al., 2017; CHELSA - Karger et al., 2017), but none of these have a resolution higher than 1 km. We suspect those global climate models underestimate variation in local climate values, even after consideration of the coarser scale. Local studies of Hong Kong meteorology have largely focused on characterizing and mitigating the effects of urbanization (e.g. Shi et al., 2018; Wang et al., 2017; Nichol et al., 2014; Liu and Zhang, 2011; Ng, 2009; Giridharan et al., 2004). Unfortunately, it appears the climate of Hong Kong's landscape as a whole has been given little notice, and we are unaware of long-term averaged climate rasters available for the region. Relevant studies that do exist include limited variables, and the data appear to be publicly unavailable. We are additionally unaware of Hong Kong data publicly available for vegetation indices such as NDVI, or topographic data other than elevation.

Therefore Hong Kong is in dire need of a comprehensive suite of accessible environmental GIS data, at a resolution finer than 1 km, suitable for species distribution modeling and other local applications. To this end, we developed new, 30 m resolution rasters of topography, NDVI, and 10 interpolated climate variables for each month of the year.

3 Methods

All data manipulation and geographic analyses were conducted in the R statistical computing environment (v3.3.2, R Core Team, 2016) using RStudio (v1.0.136, RStudio Team, 2015) unless otherwise noted. Analyses are divided into three broad categories of data products, detailed in the sections below: topographic variables, climate variables, and remote sensing variables. The variables developed were selected based on their utility in environmental research, especially SDM, as well as the availability of appropriate source data. An overview schematic of the data workflow is available in Figure S1

3.1 Topographic variables

Data on the physical characteristics of Hong Kong's landmass were assembled from remote sensing inputs, crowdsourced coastline polygons, and a digital terrain model. The topographic variables developed are coastline, elevation, slope, aspect, terrain roughness, relative elevation, distance to coast, water proximity, and urbanicity.

3.1.1 Coastline

As reclamation of land from the ocean in Hong Kong is ongoing, obtaining current data for the coastline can be challenging. Natural coastline and reservoir vectors were downloaded from OpenStreetMap (2018) and merged in QGIS (v3.01, QGIS Development Team, 2018) to produce a shapefile of polygons representing Hong Kong land area as of January 2018. All output rasters were masked to this area.

3.1.2 Elevation, slope, aspect, and roughness

A 5 m resolution Hong Kong digital terrain model (Lands Department, 2017) was upscaled using bilinear resampling. The resulting 30 m DEM was used as the elevation data throughout the study. Four other topographic predictor layers were derived directly from this DEM: aspect, slope, aspect*slope, and a roughness index. These were calculated using the Hong Kong elevation raster with the terrain() function in the R raster package, using all 8 neighboring cells (queen case). Aspect was transformed from degrees to a measure of north-south exposure ("northness") by cos(aspect*pi/180).

3.1.3 Relative elevation

Relative elevation is a measure of the difference in elevation between the pixel of interest, and the lowest pixel within a given radius. A pixel on a mountain peak has a high relative elevation, while a pixel on a flat plain has a relative elevation of 0 (regardless of its elevation above or below sea level). A set of relative elevation layers for Hong Kong were calculated at multiple scales, following the moving window approach of Bennie et al. (2010). The radii used were 60 m, 120 m, 240 m, 480 m, and 960 m. These layers are expected to be most applicable as measures of surface water drainage and therefore soil moisture as well. Relative elevation has been used as a covariate in climate interpolation as a proxy for cool air draining (Bennie et al., 2010; Ashcroft and Gollan, 2012), but were not included here as a predictor as Hong Kong lacks large valleys and other sheltered areas where this effect would be most relevant.

3.1.4 Distance to coast and water proximity

Water bodies adjacent to land areas can act as temperature buffers, contribute to evaporative cooling (Lookingbill and Urban, 2003), and influence precipitation patterns (Heiblum et al., 2011; Paiva et al., 2011); therefore considering their presence is important for climatic predictions. Here, two different methods were used to quantify water body distribution in Hong Kong: distance to coast and water proximity. A distance to coast raster, measured in meters, was produced using the distance() function in the raster package with the Hong Kong coastline shapefile described in section 3.1.1. Distance to coast did not incorporate inland water bodies. Second, water proximity (including inland water bodies) was calculated as the percent surface land in the area surrounding a given pixel. A value of 1 means that the area within a given radius is entirely terrestrial, while 0 indicates it is entirely aquatic. Multiple water proximity rasters were calculated with varying radii using a circular moving window approach like that described

by Aalto et al. (2017), to represent buffering processes at different scales. The radii used were 0.75 km, 1.5 km, 3 km, 6 km, and 12 km.

3.1.5 Urbanicity

Urbanicity rasters were developed because in densely constructed areas, urban heat island effects are expected to influence temperatures (Nichol et al., 2013; Shi et al., 2018), and therefore urbanicity may be an important predictor in climate interpolation. High rise buildings can influence temperature by blocking wind, creating shade, acting as heat sinks, and producing thermal pollution. These effects are particularly relevant for this study, as some of Hong Kong's weather observation stations are adjacent to or inside urban centers. To quantify the distribution of developed area, we used a 30 m resolution dataset of percent impervious surface (Brown de Colstoun et al., 2017), which we expect to strongly correlate with urban development. For use in climate predictions this data was smoothed using a Gaussian moving window because bulk air temperature is not expected to vary at a granular (30 m) scale. at three buffer scales (sigma = 10, 50, 100), using the focalWeight() and focal() functions in the raster R package, where type = 'Gauss'. The resulting 'urbanicity' layers were later used as climate predictors. In these rasters, completely impervious locations have a value of 100, while vegetated areas are 0.

3.2 Climate variables

Climate interpolators are often faced with the challenge of estimating climate parameters over a large area using sparse weather station observations, at least in part of the region considered (e.g. Hu et al., 2016). In contrast, interpolation in Hong Kong is benefitted by a relatively small geographic area and a quite dense network of weather data provided by dozens of permanent weather stations (Hong Kong Observatory, 2018; see Figure S2). Here we use multiple linear regression to predict geographic climate patterns using weather station training points and raster covariates. This is followed by thin plate spline (TPS) interpolation (see Wahba, 1979) of the regression model residuals. TPS is a widely used approach in climate interpolation (e.g. New et al., 2002; Fick and Hijmans, 2017), which fits a curved surface to irregularly distributed points. This two-step interpolation (regression followed by TPS) was based on the approach of Meineri and Hylander (2017).

Weather station observation data and geographic coordinates were downloaded from the web portal of the Hong Kong Observatory (2018). As the goal was to produce a representation of long-term but modern climate, measurements over 20 years (1998 to 2017) were included. To ensure averages were reliable, weather stations were only included for interpolation of each variable if at least 8 years of complete data were available within the 20 year window. The minimum number of stations used for each model is provided in Table 2. Monthly observations of ten variables were obtained: maximum temperature, mean daily maximum temperature, mean daily temperature, mean daily minimum temperature, minimum temperature, mean dew point, mean relative humidity, mean wind speed, mean air pressure, and total rainfall.

Climate interpolation consisted of two main steps. First, a linear model was built for each climate variable for each month of the year. Independent variables were selected by searching the literature for similar studies, and choosing predictors we expected to have an influence on climate at this regional scale. When necessary, each predictor was statistically transformed to approach a normal distribution. The six topographic predictors used as model building candidates were: elevation, log-transformed distance to coast, exponentially transformed fine and coarse water proximity, log-transformed urbanicity (sigma = 50), and 'northness' - the cross product of aspect and slope. The water proximity layers were products of additively combining multiple scale rasters into fewer predictors: fine water proximity was the sum of 0.75 km, 1.5 km, 3 km scale rasters, while coarse was the sum of 6 km, and 12 km. The six model predictors were tested for collinearity and no problems were found. Linear models were built using the lm() R function. All predictors were initially included, then using the step() function, pared down in each regression model using stepwise bidirectional selection based on AIC, using 4 degrees of freedom as a penalty to make predictor selection stricter than the default. The resulting regression model was used to calculate a climate value at each grid cell based on a linear relationship with the selected predictors.

Second, to adjust for local variation in climate that is not associated with topography, the linear model residuals at each station were calculated and interpolated using the thin plate spline approach implemented in the fields R package. The lambda smoothing parameter, which determines how closely the fitted surface matches input values, was set to 0.01. This low lambda value was selected because of the relatively high confidence in the long-term averaged weather station values (based on at least 8 years of data). This effectively produces a smoothed layer of local deviation from the linear model, which was used to additively adjust the results of the linear model predictions and produce finalized climate rasters.

We measured the spatial predictive ability of models using ten-fold cross-validation (Dobesch et al., 2007). In each validation round, 10% of weather stations were reserved as a test dataset and the remainder were used for training. Average root mean squared error of the test data subset from the final model prediction was used as an error measurement. To normalize these error measures across the climate variables, we adjusted them as a percentage of the standard deviation of the initial weather station values measured. This cross-validation procedure was used only to produce these validation measurements. The finalized monthly climate rasters described above were trained using all available data.

The finalized monthly rasters were then summarized into raster layers that characterize yearly climatic means and variation. These include 19 "bioclimatic" variables using the biovars() function in the dismo R package (Hijmans et al., 2017), which are specifically suited for species distribution modeling and other ecological purposes. This also allows our data to be compared with other climate data products that use the same calculations. Because those calculations only use rainfall and average daily maximum and minimum temperatures in each month, we also produced yearly average layers of dewpoint, relative humidity, mean daily temperature, air pressure, and wind speed. Also provided are layers of highest and lowest average monthly

extreme temperatures, and their difference (extreme temperature annual range). These two variables characterize temperature extremes experienced in a given location better than the bioclimatic variables.

For comparison with global climate data products, we resampled bioclimatic variables to the same (1 km) resolution as WorldClim using bilinear interpolation. Only pixels present in both data products were used for comparisons.

3.3 Remote sensing data

Normalized difference vegetation index (NDVI) is a common metric of vegetation presence and density derived from satellite imagery. To calculate NDVI, Landsat 8 images (U.S. Geological Survey, 2018) of Hong Kong were obtained. We downloaded one image from March 2016 that covers much of Hong Kong except for the far eastern areas, and is free of clouds. This was supplemented with an image from March 2018 after adjustment, so that all land areas of the region were included. NDVI calculations were completed using the standard equation:

NDVI = (NIR−Red)/(NIR+Red)

Where NIR is near-infrared (Landsat band 5: 0.851 to 0.879 μm) and Red is visible red radiation (Landsat band 4: 0.636 to 0.673 μm). The resulting NDVI value varies between 1 and -1, where higher values correspond with denser vegetation.

4 Results and discussion

Results of this environmental analysis of Hong Kong include 48 rasters and one vector file. All rasters are provided at an identical 1 arc second (30 m) resolution and in the WGS84 geographic coordinate system. Summary values and filenames are provided in the data repository.

4.1 Topographic variables

Distance to coast results show that approximately 42% of Hong Kong's land area is within 1 km of the coastline. However it is apparent that inland areas often feature steep inclines, as half of Hong Kong's land is above 84 m elevation.

For variables like relative elevation, urbanicity, and water proximity, the ideal scale of raster calculation is dependent on the desired effect to be captured, and perhaps other characteristics of the landscape in question. For this reason, we provide these rasters calculated at multiple buffer scales.

Urbanicity results show that the majority of land in Hong Kong is not near urban areas, as the median raster values are below 4% urban at all scales calculated (Table 1). This shows that although Hong Kong has extremely dense urban cores, most of its mountainous terrain is unpopulated.

4.2 Climate variables

Minimally, 32,024 monthly weather station measurements over 20 years (1998 to 2017) were used to construct climate models for all months and variables at finer resolution compared to global datasets (Fig. 2). High weather station density and availability of data on multiple candidate topographic climate-forcing factors allowed for high confidence in many climate variable models, especially those related to temperature (Figs. 3, 4). The climate interpolation results include monthly models of ten variables including temperature, precipitation, and humidity, making a total of 120 individual models produced (monthly models of three temperature variables are shown in Fig. 5). For all variables, the predictors included in monthly models are displayed in Figure 6, and the number of stations with data included is in Table 2.

4.2.1 Temperature

Temperature was found to vary considerably across Hong Kong, with more than 6°C difference in mean annual temperature between the highest mountain peaks (>900 m, <18°C) and some low-lying urbanized areas (>24°C). While mean and minimum temperature are highest in urban areas, maximum temperature shows a different pattern with a maximum in inland valleys in the northern New Territories. This pattern may be explained by urban heat retention: buildings act as heat sinks which absorb solar radiation during the day, and slowly release heat at night, causing increased minimum temperatures (see Oke, 1982). The high maximum temperatures in inland valleys may be due to reduced air circulation in sheltered locations, and lack of complex vegetation or urban structures providing shade. The high accuracy of temperature models (Figs. 3, 4) is likely due to a strong association with elevation; elevation was by far the most commonly included predictor for temperature models (Fig. 6). Urbanicity was important for mean and minimum temperature, but not maximum temperature. Water proximity and coast distance were differentially included depending on the variable, while aspect*slope rarely had an effect.

4.2.2 Rainfall

In our models, the highest annual rainfall (bio12) areas in Hong Kong (>2500 mm annually) are inland and at high elevations, presumably because of condensation from humid air as it passes over mountains. Areas near the coast, particularly small outlying islands and the eastern coast in Lung Kwu Tan receive the lowest amount of annual rainfall (<1600 mm). Precipitation of driest month (bio14) was uniformly low, ranging from 20 to 40 mm, but the relative pattern of high and low precipitation areas remained similar. The most commonly included model predictor was fine-scale water proximity (Figure 6). Elevation was predictive for 5 out of 12 months, but few other topographic predictors were useful. Seasonality of rainfall in Hong Kong is strong. Averaged across all locations, 52% of total yearly rainfall was recorded in three months (June through August). Rainfall models were informed by more weather stations than any other climate variable (Table 2), but they have the highest relative standard error (Fig. 3) and therefore the lowest accuracy. Because they are influenced by both global and locally variable wind patterns, precipitation distributions are notoriously difficult to predict, especially in urban areas (Cristiano et al., 2017).

Deleted: A few special notes on the limitations of predicted temperature values in dense urban areas: temperature predictions of our landscape-level climate models will not reflect the high spatial variation in temperature found in urban microclimates. Although the manned Kowloon HKO weather station is inside a densely populated area, as pointed out by Nichol and To (2012) it is still in a small parklike area surrounded by trees, and therefore is not representative of the most densely urbanized areas of Hong Kong. Other stations in urban areas are similarly near green spaces or otherwise open areas. Higher resolution (say 5 m or 1 m) studies of urban thermal distributions would strongly benefit from analysis of wind patterns, building height, thermal pollution, and other factors (e.g. Shi et al., 2018). Therefore granular, ground-level temperatures in urban areas are likely substantially different than the broader air temperature values our models provide. One area of particular interest is the Hong Kong Airport, a massive area reclaimed from the ocean, north of Lantau Island. The weather station here has the highest urbanicity value, because the airport is mostly impervious surface. However it lacks properties of truly urbanized areas, with no permanent population lacking typical urban morphology. Therefore climate variables in this area may be biased, especially for variables like wind speed that would be affected by the presence of tall buildings. The airport weather station often had the highest mean temperatures recorded, perhaps indicating that extensive impervious surface is more important than wind blockage or thermal pollution for maintaining high temperatures.

Moved (insertion) [1]

Deleted: Although r

Deleted: Our relatively poor results may be explained by this, as well as lack of appropriate local predictors.

4.2.3 Dew point, humidity, pressure, and wind speed

Dew point exhibits a similar pattern to other temperature variables, with mean annual dew point ranging from 15.5°C at 5 mountain peaks to around 19°C on small islands and lower areas. Mean annual relative humidity reaches a maximum of about 90% at Tai Mo Shan, while many urban areas in Kowloon, Tuen Mun, and Yuen Long are between 70 and 75%. Surprisingly, mean annual air pressure has a positive correlation with elevation; the highest values (reaching 1014 hPa) are at mountain peaks, and particularly low values (as low as 1012.5 hPa) in coastal areas of southern and western Hong Kong. Mean annual wind speed is also strongly associated with elevation, with mean annual values above 30 km/h on Lantau Island mountain peaks, down to below 5 km/h in interior low elevation areas of the New Territories.

4.2.4 Comparisons with global climate data

Our new climate models are compared with a recent global climate dataset to identify differences in predictions of Hong Kong climate values (Fig. 7). WorldClim 2 was produced using a similar interpolation approach with regression modeling and thin plate spline interpolation, but also included satellite-derived covariates in addition to topography (Fick and Hijmans, 2017). Because WorldClim incorporates vast amounts of data from multiple databases covering overlapping geographic and political entities, it is difficult to ascertain exactly which individual weather stations were included, and we were unable to determine whether any Hong Kong weather stations were included or if the datasets are completely independent. However, the model predictions differ substantially (Figs. 1, 7; Table 3). Our models generally indicate greater spatial variation than WorldClim, with cool areas colder, warm areas hotter, and wet areas wetter. For example in average low temperature of coldest month (bio6), high elevation areas could be more than 2°C lower, and urban areas more than 2°C higher than WorldClim indicates (Fig. 7a). To further quantify differences in values between these two datasets, for each of the 19 bioclimatic variables we calculated the standard deviation of raster values (Table 3). All of our interpolated climate rasters had a higher standard deviation than their WorldClim 2 counterparts. These results suggest that unless global climate models increase in resolution and accuracy, regional models will remain critical for local applications.

4.3 Remote sensing variable

The NDVI data represents vegetation quality and density based on two merged satellite images, both in March of their respective years. Although this is only an instantaneous representation of NDVI, we expect it to correlate strongly with the spatial pattern of vegetation density throughout the year. Certain plant species shed and regenerate their leaves during specific months ranging from winter through mid-summer, but Hong Kong's woody vegetation is overall evergreen (Dudgeon and Corlett, 1994), so seasonal changes in NDVI are not expected to be drastic. NDVI values above 0.4 include Hong Kong's densest forests, while unvegetated or urbanized areas are well below 0.1. The densest vegetation (> 0.4 NDVI) in Hong Kong tends to be on slopes

between 100 m and 400 m elevation (Fig. 8), and is distributed between Hong Kong Island, Lantau Island, and the New Territories. One exception is the verdant mangrove forests, at sea level. The patchy distribution of high density vegetation likely reflects the effects of historical deforestation. The largest patches are found on the southeastern slopes of Tai To Yan in the New Territories. The relative distribution of NDVI classes along Hong Kong's elevational gradient is shown in Figure 8. Future work could determine to what extent NDVI changes over time, in response to seasonality or recent weather. The limiting factor is the availability of data of adequate temporal resolution, as many satellite images of Hong Kong are obscured by cloud cover or degraded by poor air quality.

4.4 Value and Utility

This new data will benefit environmental research, and specifically SDM studies, in two main ways. First, it will enable finer scale analyses than previously possible. For SDM, this means improved detection of climatic microrefugia (Meineri and Hylander, 2017), and the ability to differentiate between human altered habitat and natural areas. Rampant development and a shifting climate make this knowledge of local species persistence more important than ever. Additionally, this is especially relevant in Hong Kong where topography varies dramatically, and where urban areas form a complex mosaic with undeveloped expanses.

Second, we provide a diverse array rasters derived from multiple independent data sources, but in a single resolution and format to facilitate further analysis and synthesis of meaning. For SDM, these diverse layers have distinct advantages over datasets that only contain climate data. Compared to climate data alone, using diverse predictors including topographic characteristics have been shown to be important variables for accurate SDM results, such as predicting the spread of invasive species in new ranges (Peterson and Nakazasa, 2008). However benefits of non-climate data may only be evident in finer scale SDMs (Luoto et al., 2007).

Finally, such high quality, diverse geographic data is especially uncommon in tropical regions, where improved knowledge for environmental research and biological conservation is most needed. According to Rapoport's Rule, tropical species are more likely to have smaller distributions (Stevens, 1989), and therefore future execution of local SDM studies to understand their ranges are particularly important.

4.5 Limitations and next steps

Here we outline how shortfalls of the data presented may be improved in the future. First, though we inferred Hong Kong's pattern of urban development from impervious surface data, this is less than ideal because in addition to concrete, bare soil or rock are sensed as impervious. Also, it cannot differentiate dense urban cores of high-rises from large paved areas. For climate modeling, an urbanicity measure that considers building height or population density at a 30 m or finer scale could be preferable.

Second, while our temperature rasters should accurately represent air temperature in open areas, they do not reflect the high spatial variation in temperature found in urban microclimates. For

example, although the manned Kowloon HKO weather station is inside a densely populated area, as pointed out by Nichol and To (2012) it is still in a small parklike area surrounded by trees, and therefore is not representative of the most densely urbanized areas of Hong Kong. Other stations in urban areas are similarly near green spaces or otherwise open areas. Higher resolution (say 5 m or 1 m) studies of urban thermal distributions would strongly benefit from analysis of wind patterns, building height, thermal pollution, and other factors (e.g. Shi et al., 2018). Therefore granular, ground-level temperatures in urban areas are likely substantially different than the broader air temperature values our models provide.

Similar to other climate interpolation studies, bias in the physical locations of automatic weather stations may be of concern. Weather stations are often intentionally placed in flat, open areas with the goal of measuring weather that is relevant to a broad geographic area, rather than locations that may experience unique local climate. It may be for this reason that Slope*Aspect was infrequently useful for model construction, as few stations are on steep slopes. Elevational distribution of stations may also be a source of bias; although a weather station operates at the highest point in Hong Kong (Tai Mo Shan, 955 m), there are only two other stations above 600 m.

Finally, while we used cross-validation to measure the spatial predictive ability of the climate models, this method is only able to test models against locations where weather stations are present; validation based on an independently collected dataset would be ideal. One common validation method is to use weather data loggers placed across elevational and land-use gradients (Meineri and Hylander, 2017). Such an approach would allow for explicit testing and comparing predictiveness of climate products for different areas of Hong Kong.

Important gaps in Hong Kong geographic data remain. Models projecting future climate scenarios would enable biodiversity change predictions, with additional variables like cloud cover and solar radiation useful. A discrete classification of habitat type would be useful for ecological research, and quality soil type data is lacking. Availability of such data for Hong Kong would complement the findings of this project, which significantly advance our understanding of geographic heterogeneity in this complex tropical region.

Conclusions

This diverse set of 30 m resolution topography, climate, and remote sensing data include the first published interpolation of long-term climate averages specific to Hong Kong. Our findings suggest that global interpolated climate datasets are limited by their resolution, and underestimate local climate variability. Therefore the availability of such local data will remain critically important for the forseeable future. This new data will allow for a new generation of studies in Hong Kong, and enable connections between environmental data and biotic patterns at a much finer scale than previously possible. Aside from clear uses in conservation, ecological and biogeographic research, we also expect this freely accessible dataset to be broadly applicable for many sectors, including tourism, hydrology, recreation, agriculture, mapmaking, and real estate.

Deleted: One area of particular interest is the Hong Kong Airport, a massive area reclaimed from the ocean, north of Lantau Island. The weather station here has the highest urbanicity value, because the airport is mostly impervious surface. However it lacks properties of truly urbanized areas, with no permanent population lacking typical urban morphology. Therefore climate variables in this area may be biased, especially for variables like wind speed that would be affected by the presence of tall buildings. The airport weather station often had the highest mean temperatures recorded, perhaps indicating that extensive impervious surface is more important than wind blockage or thermal pollution for maintaining high temperatures.

Deleted: .

Deleted: A

Deleted: W

Deleted: . However

Deleted: ,

Deleted: and

Deleted: Finer grain, microclimatic variables (Lembrechts et al., 2018) based on remote sensing data and temperature loggers could be a next step for local climate modeling.

Deleted: and

Deleted: would be

Deleted: .

Deleted: A series of studies in the past, commissioned by the Hong Kong government, classified Hong Kong's terrestrial habitats (e.g. Environmental Resources Management, 2009), however the results of these very thorough remote sensing studies are not accessible. Qu

Deleted: , and as mentioned before a better index of population density or built-up area would be welcomed

Deleted: results

Deleted: and

Deleted: benefit geographic studies in Hong Kong. .

6 Data availability

GeoTIFF raster and shapefile documents can be downloaded from figshare: https://doi.org/10.6084/m9.figshare.6791276. A document in the repository includes file names, descriptions, and summary statistics for all provided rasters. Individual monthly rasters for each of the 10 climate variables are available as a compressed zip file.

Citations Added:

Luoto, M., Virkkala, R., and Heikkinen R. K.: The role of land cover in bioclimatic models depends on spatial resolution. Global Ecol. Biogeogr., 16, 34-42, 2007.

New, M., Lister, D., Hulme, M., and Makin, I.: A high-resolution data set of surface climate over global land areas. Clim. Res., 21, 1-24, 2002.

Oke, T. R.: The energetic basis of the urban heat island. Quart. J. R. Met. Soc., 108, 1-24, 1982.

Peterson, A. T., and Nakazawa, Y.: Environmental data sets matter in ecological niche modelling: an example with Solenopsis invicta and Solenopsis richteri. Global Ecol. Biogeogr., 17, 135-144, 2008.

Stevens, G.: The latitudinal gradient in geographical range: how so many species coexist in the tropics. Amer. Nat., 133, 240-256, 1989.

Wahba, G.: How to Smooth Curves and Surfaces with Splines and Cross-Validation. Ft. Belvoir: Defense Technical Information Center, 1979.

Figures and Appendix Added:

Deleted: https://figshare.com/s/3a5634e36e80dc33444c (or)

Deleted: (once data is published). The readme

Deleted: from the authors upon request

Moved (insertion) [3]

Moved up [3]: Luoto, M., Virkkala, R., and Heikkinen R. K.: The role of land cover in bioclimatic models depends on spatial resolution. Global Ecol. Biogeogr., 16, 34-42, 2007.

Moved down [4]: Peterson, A. T., and Nakazawa, Y.: Environmental data sets matter in ecological niche modelling: an example with Solenopsis invicta and Solenopsis richteri. Global Ecol. Biogeogr., 17, 135-144, 2008.
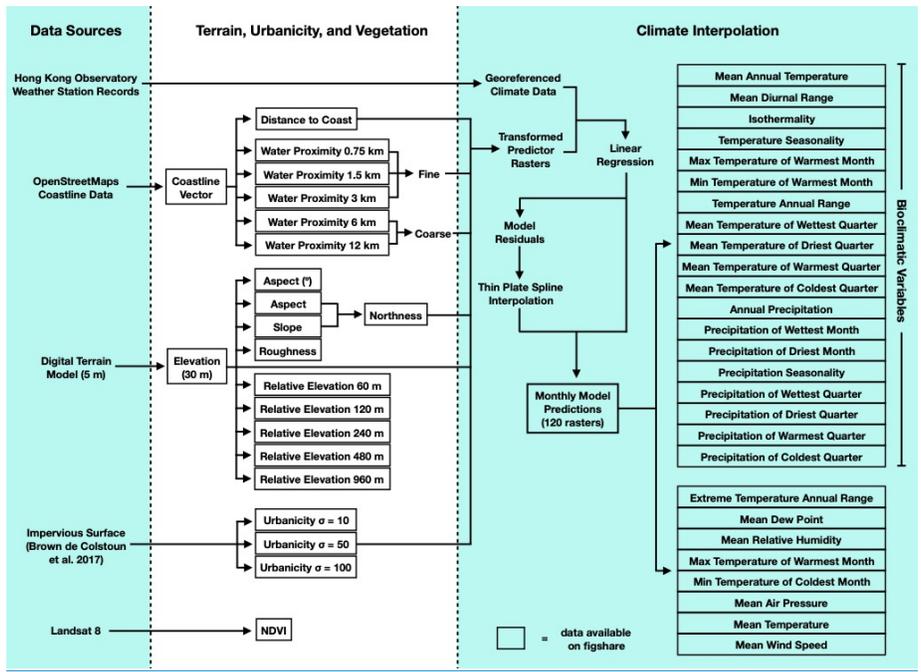
Moved (insertion) [4]

Figure S1. Schematic of data products and the sources that informed them. Items enclosed in a box represent the files available for download from the figshare repository.
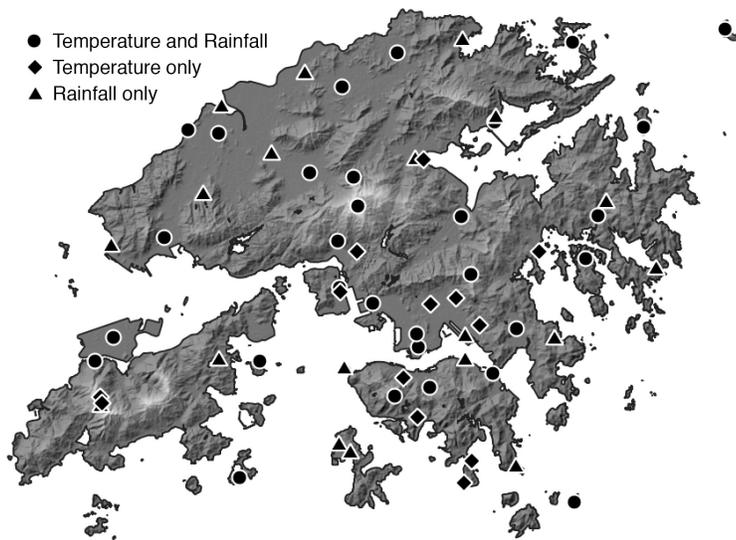
Figure S2. Permanent weather stations operated by the Hong Kong Observatory. Symbols indicate what type of data is available from each station: temperature, rainfall, or both.

Appendix 1. Glossary of variable definitions

Maximum temperature	the highest temperature observed within a month
Mean daily maximum temperature	the mean of all daily high temperatures within a month
Mean daily temperature	the mean of all temperatures within a month
Mean daily minimum temperature	the mean of all daily low temperatures within a month
Minimum temperature	the lowest temperature observed within a month
Mean dew point	the mean of all dew point observations within a month
Mean relative humidity	the mean of all relative humidity observations within a month
Mean wind speed	the mean of all wind speed observations within a month
Mean air pressure	the mean of all air pressure observations within a month
Rainfall	the total of all rain recorded within a month
Relative elevation	the difference in elevation between the pixel of interest, and the lowest pixel within a given radius
Distance to coast	geometric distance between the pixel of interest and the nearest oceanic coastline
Water proximity	percent of area that is terrestrial within a given radius of the pixel of interest
NDVI	Normalized Difference Vegetation Index
Urbanicity	measure of area that is impervious surface within a given radius of the pixel of interest