Global patterns of fern species diversity: An evaluation of fern data in GBIF

Hong Qian a,*, Jian Zhang b, Mei-Chen Jiang b

a Research and Collections Center, Illinois State Museum, 1011 East Ash Street, Springfield, IL 62703, USA
b Zhejiang Tiantong Forest Ecosystem National Observation and Research Station, School of Ecological and Environmental Sciences, East China Normal University, 200241 Shanghai, China

A R T I C L E   I N F O
Article history:
Received 7 July 2021
Received in revised form 11 October 2021
Accepted 13 October 2021
Available online 27 October 2021

Keywords:
Climate
Data bias
Fern
GBIF
Species diversity
Species list

A B S T R A C T
Despite that several studies have shown that data derived from species lists generated from distribution occurrence records in the Global Biodiversity Information Facility (GBIF) are not appropriate for those ecological and biogeographic studies that require high sampling completeness, because species lists derived from GBIF are generally very incomplete, Suissa et al. (2021) generated fern species lists based on data with GBIF for 100 km × 100 km grid cells across the world, and used the data to determine fern diversity hotspots and species richness–climate relationships. We conduct an evaluation on the completeness of fern species lists derived from GBIF at the grid–cell scale and at a larger spatial scale, and determine whether fern data derived from GBIF are appropriate for studies on the relations of species composition and richness with climatic variables. We show that species sampling completeness of GBIF is low (<40%) for most of the grid cells examined, and such low sampling completeness can substantially bias the investigation of geographic and ecological patterns of species diversity and the identification of diversity hotspots. We conclude that fern species lists derived from GBIF are generally very incomplete across a wide range of spatial scales, and are not appropriate for studies that require data derived from species lists in high completeness. We present a map showing global patterns of fern species diversity based on complete or nearly complete regional fern species lists.

Copyright © 2021 Kunming Institute of Botany, Chinese Academy of Sciences. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co., Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction
Ferns, which include about 12,000 species worldwide (Hassler, 2004–2021), are one of the oldest and the most species-rich groups of vascular plants (Mabberley, 2008; Qian et al., 2021a). Fern propagules are spores, which are small (usually <0.1 mm in equatorial axis and polar axis; Adsersen, 1995), and are capable of dispersing thousands of kilometers by wind (Wolf et al., 2001). Ferns are generally distributed broadly, and fern distributions are thought to be more in equilibrium with climate than most other groups of vascular plants (Qian, 2009). Fern species richness exhibits marked variation among areas across the globe (Weigand et al., 2020), which is thought to be driven by environmental factors (Kreft et al., 2010; Rhine et al., 2019). Therefore, ferns are an ideal group of vascular plants for the study of geographic and ecological patterns and drivers of plant diversity at global and regional scales.

In an article by Suissa et al. (2021), the authors analyzed species occurrence data downloaded from the Global Biodiversity Information Facility (GBIF) to address several questions. They converted the geo-referenced occurrences downloaded from GBIF to fern species lists for each of 100 km × 100 km grid cells across the globe, explored geographic patterns of fern species richness and endemism, and tested for correlations between individual environmental variables and richness and speciation rate. The robustness of the conclusions of their study heavily depends on the quality of the data used in their study (i.e. the completeness of fern species list for each of the 100 km × 100 km grid cells).

While GBIF occurrence data are useful for biological conservation and some ecological and biogeographic studies, it may not be appropriate to use the data in those studies that depend on species lists derived from the GBIF data. This is because previous studies have shown that species lists derived from GBIF occurrence data are
commonly very incomplete and the completeness of a species list varies non-randomly across the globe. For example, Yesson et al. (2007) found that GBIF included only 31% of global Fabaceae species richness while large parts of the world are data deficient. Beck et al. (2013) found that the GBIF data for European moths provided less information on species' geographic ranges and climatic niches than an independent data compilation based on museum collections and published literature. Qian et al. (2018) showed that the completeness of species lists of vascular plants derived from GBIF is only 13% for the Chinese counties examined in their study (8195 km² per county on average) and is only 37% for the Chinese provinces (342,749 km² per province on average). Qian et al. (2018) also showed that the relationships between species richness and climate can be substantially biased when species richness derived from GBIF is used in an analysis relating species richness to climatic variables. Recently, Qian et al. (2021b) showed that species lists of vascular plants derived GBIF for 100 km × 100 km grid cells in Africa account for less than 37% of the species in their full species lists.

Suissa et al. (2021) included only two third of the fern species worldwide in their study. Because the completeness of species lists derived from GBIF data decreases with decreasing spatial scale, as shown in Qian et al. (2018, 2021b), the completeness of the fern species lists for the 100 km × 100 km grid cells used in their study are likely very low, at least in some regions. Suissa et al. (2021) pointed out that “misidentified records or those with problematic localities can bias biodiversity analyses”, but they overlooked the problem of using substantially incomplete species lists in their study. To determine whether the conclusions of their study are valid, it is necessary to evaluate the quality of the data used in their study. Here, we report an evaluation on their data.

2. Materials and methods

Suissa et al.’s (2021) primary analyses were based on species lists at the spatial resolution of 100 km × 100 km, but their study also invoked larger geographic areas, such as those regions that were used to characterize regional patterns and hotspots of species richness and endemism. Accordingly, we evaluate the completeness of fern species lists derived from the GBIF occurrence records reported in Suissa et al. (2021) for geographic areas at both regional scale and 100 km × 100 km grid cell scale. At a regional scale, we extracted fern occurrence records from World Plants (WP; https://www.worldplants.de/) and Plants of the World online (POWO; http://www.plantsoftheworldonline.org/) for geographic units mostly defined in Brummitt (2001), which are geographic units at the level 3 in most cases, and are geographic units at the level 4 defined in Brummitt (2001) for several countries with large extents of latitude or longitude or both. These countries are: Argentina, Australia, Brazil, Canada, Chile, China, Mexico, Russia, South Africa, and USA. For Russia, geographic units located in Europe are political regions shown in Map 5 of Brummitt (2001), and geographic units located in Siberia and Russian Far East are those shown in figure 1 of Zhang et al. (2018). A total of 470 geographic regions, as shown in Figure 1, were used to document global fern distributions. Regional species lists derived from the data extracted from WP and POWO were supplemented by additional data sources, e.g. GBIF data for global fern occurrences reported in Suissa et al. (2021), Flora of China online (http://www.efloras.org/flora_page.aspx?flora_id=2), and PLANTS Database (https://plants.usda.gov/home). Of the 470 geographic regions, 457 had at least one fern species, and were analyzed in this study. These regional species lists, which were considered as ‘complete’ or ‘nearly complete’ species lists for the regions, were compared with those regional species lists derived solely from the GBIF occurrence records reported in Suissa et al. (2021). For a particular geographic region, we determined the completeness of the species list derived solely from the GBIF data by dividing the number of species in the GBIF-based species list by the number of species in the ‘complete’ species list of the region, as described above. Botanical nomenclature for ferns from all the above-mentioned data sources was standardized according to Hassler (2004–2021), which was also used to standardize fern nomenclature in Suissa et al. (2021).

Complete fern species lists for 100 km × 100 km grid cells are generally not available, and cannot be generated for the vast majority of the global land surface due to lack of sufficient small-scale complete species lists. However, because the completeness of species lists of vascular plants for counties in USA are high, particularly for those counties which, or parts of which, have been botanized with an aim of compiling their complete species lists (appendix A of Qian et al., 2007), and because county-level species lists of vascular plants have been used to address species richness questions in previous studies (e.g. Stohlgren et al., 2003), we divided the contiguous USA (with 48 states) into 100 km × 100 km grids and used county-level distributions available at the PLANTS Database (https://plants.usda.gov/home) and local (nature reserve or park) plant checklists published in Weiser et al. (2018) to generate fern species lists for 100 km × 100 km grid cells in USA. Similarly, county-level plant distributions in China have been used to generate species lists for 100 km × 100 km grid cells in China, which were used in studies on species richness patterns (e.g. Feng et al., 2016). We divided China into 100 km × 100 km grid cells, and used county-level fern distributions and local (nature reserve and park) species lists available online (e.g. the National Specimen Information Infrastructure, www.nsii.org.cn/) or the literature (e.g. Qian et al., 2018) to generate fern species lists for 100 km × 100 km grid cells in China. The approach that we used to generate species lists for grid cells based on county-level and local species lists has been commonly used in the literature (e.g. Feng et al., 2016). GBIF occurrence records reported in Suissa et al. (2021) were also used when we generated fern species lists for 100 km × 100 km grid cells. For both China and USA, we only used those 100 km × 100 km grid cells which have complete fern species lists for counties or localities within each of them, based on the county-level or local floras used in Qian et al. (2018) for China and Weiser et al. (2018) for USA. As a result, we included 267 grid cells in our analysis (115 in China, 152 in USA). For both countries, a Mollweide (equal-area) projection was used to divide them into 100 km × 100 km grid cells. Botanical nomenclature in each data set was standardized according to Hassler (2004–2021).

One of the key components of Suissa et al.’s (2021) study was to analyze species richness and hotspots in climate spaces, particularly in a temperature–precipitation space (e.g. their figure 5), using regression models. Although Qian et al. (2018) showed that the relationships between species richness and a given climatic variable can differ not only in strength but also in direction between the regression models based on data derived from complete species lists and those based on the data derived from GBIF for vascular plants, it is not clear whether this conclusion applies to ferns because the relationships between species richness and climatic variables differ substantially between pteridophytes, the vast majority of which are ferns, and seed plants (Kreft et al., 2010). Accordingly, we also assessed whether using data derived from incomplete GBIF fern species lists in regression models would significantly affect the results on the relationships between species richness and climatic variables. Mean annual temperature, annual precipitation, minimum temperature of the coldest month, precipitation during the driest month, temperature seasonality, and precipitation seasonality represent the mean, extreme and variability of temperature and precipitation. Because these climatic...
variables are commonly included in studies on geographic and ecological patterns of plant diversity (e.g. Kooyman et al., 2012; Weigelt et al., 2015; Qian et al., 2017, 2021), and some of them were also used to build climate spaces in Suissa et al. (2021), our analysis emphasized on these six climatic variables. We obtained climatic data for each 100 km × 100 km grid cell from the WorldClim database (http://worldclim.org/version2), using data at the 30-arc-second resolution. We used spatial regressions (simultaneous autoregressive (SAR) models) in our analyses, which accounted for spatial autocorrelation (Kissling and Carl, 2008). We investigated the effect of data completeness on inference from the richness–climate relationship. We ran SAR models separately for the two data sets of each country (i.e. a data set derived from complete species lists (full data set), a data set derived solely from GBIF) and compared effect sizes (standardized regression coefficient) within models and R² values between models. Species richness was transformed by log10 (x + 1). Each climatic variable of a regression was standardized to have a mean of zero and a standard deviation of one. Spatial Analysis in Macroecology (www.ecoevol.ufg.br/sam) was used to conduct SAR.

Fig. 1. Comparison between fern species density (i.e., species richness was divided by log10-transformed area in square kilometer) derived from GBIF alone and that derived from WP, POWO and additional sources for geographic regions (countries or sub-countries) across the world. (a) Species density based on GBIF, (b) species density based on WP, POWO and additional sources, (c) percentage of species richness derived from GBIF over species richness derived from WP, POWO and additional sources (i.e., completeness (%) of fern species lists derived from GBIF).
3. Results and discussion

Suissa et al. (2021) report that their cleaned version of the GBIF data for ferns includes 7865 species. Their study intends to include only binomials (i.e. species-level taxa), as indicated in their Appendix S2, but because they mistakenly treated trinomials (infraspecific taxa; e.g. Asplenium affine var. mettenii) as binomials (species), some species were counted more than once in their study. For example, the species A. affine was counted four times in their study (i.e. A. affine, A. affine var. pectin, A. affine var. pectin). With duplicate species names being removed, Suissa et al.’s study actually included 7462 species, which included 62% of fern species in the world (Hassler, 2004–2021). Our analyses reported here used the corrected version of Suissa et al.’s data set.

At the beginning of the Results section of Suissa et al. (2021), the authors state that “Species richness per grid cell ranged from 3 to 929 species.” This statement is incorrect, because many grid cells having 1–2 species, as shown in appendix S2 of Suissa et al. (2021), were ignored by the authors. Thus, Suissa et al. (2021) incorrectly presented their data in their figure 1. Because many grid cells, particularly those located in arid regions, truly have 1 or 2 fern species, these grid cells should be shown in their figure 1. We have updated their figure 1 by using a corrected version of Suissa et al.’s data set, as mentioned above (i.e. combining trinomials with their respective binomials), and adding grid cells with 1 or 2 species on the map (Fig. S1).

For the 457 geographic regions across the world, each of which has, on average, 294,310 km², the completeness of a fern species list derived from GBIF was, on average, only 51% of its full fern species list (Fig. 1). About 54% of the geographic regions each have less than 60% of the completeness in their species lists derived from GBIF (Fig. S2). Regional fern species lists derived from GBIF are substantially incomplete for a large geographic extent from northern Africa eastward to eastern Asia (Fig. 1c), including China and India, which are rich in fern diversity (Fig. 1b).

At the grid–cell scale (i.e. 100 km × 100 km), our analysis showed that fern species lists derived from GBIF included less than 20% of all the species in each grid cell for 92% of the 115 grid cells sampled from China (Fig. 2). Only 2% of the species lists derived from GBIF for China each had more than 40% of all the fern species in the grid cells. The completeness of species lists derived from GBIF for grid cells in the USA was higher than that for China, but fern species lists derived from GBIF included less than 60% of the species in each of the majority (53%) of the 152 grid cells sampled from the USA (Fig. 2). When complete fern species lists were considered, each grid cell had, on average, 231 species in China and 29 species in USA, i.e. actual fern species richness per grid cell in China is eight times as high as in USA. However, when fern species lists derived from GBIF were considered, fern species richness per grid cell in China is nearly the same as that in the USA (18.8 versus 18.3 species). The completeness of species lists derived from GBIF was lowest for the most species-rich grid cells (Fig. S3). Clearly, geographic patterns and hotspots of fern species diversity determined according to the GBIF data are substantially biased. Our analysis showed that the completeness of species lists derived from GBIF tended to be lower in areas with richer floras (Fig. S1), suggesting that identifying diversity hotspots solely based on data in GBIF, as did in Suissa et al. (2021), is not reliable.

This problem with Suissa et al. (2021) can be easily seen from their figure 1. For example, their figure 1 showed that most grid cells in Japan are among the grid cells with the highest fern species richness across the globe, and have much higher fern species richness than those grid cells located in the Hengduan Mountains.
and southeastern China. However, both regional floras in the literature and our data show that local and regional fern species richness in the Hengduan Mountains and southeastern China is much higher than that in Japan. For example, the five most species-rich grid cells in Japan each has 190 to 218 fern species based on data from GBIF, but many grid cells in southeastern China each have 220 to 400 fern species in the data we analyzed, and these grid cells commonly each have few to none fern species in the GBIF data set analyzed by Suissa et al. (e.g., the grid cell whose centroid is located at 26.41° N and 117.38° E has 377 fern species but it has no species in the GBIF data reported in Suissa et al.). Such cases occur in many regions across the world. For example, the fern flora of India is rich, with over 1000 species (according to WP and POWO). However, few grid cells across the entire India south of the Himalayas have fern species in figure 1 of Suissa et al. (2021). We believe that the fern diversity hotspots identified by Suissa et al. (2021) largely reflect the availability of occurrence records in GBIF, rather than true fern diversity hotspots, and many true fern diversity hotspots, such as the Hengduan Mountains, southeastern China and tropical mountains in India, have not been identified in their study.

Our analysis showed that using fern species lists derived from GBIF can substantially bias the relationships between species richness and climate not only in strength but also in direction (Table 1). For example, for grid cells in China, when species richness derived from complete species lists was used in a regression analysis with six climatic variables being included as independent variables, the model explained 65.5% of the variation in fern richness and annual precipitation was the strongest correlate of fern richness and was positively associated with fern richness (Table 1), which is consistent with findings reported in previous studies (e.g., Nagalingum et al., 2015). By contrast, when species richness in the model was replaced by that derived from the GBIF data reported in Suissa et al. (2021), the model explained only 34.2% of the variation in fern richness and annual precipitation was not only the weakest correlate of fern richness but also was negatively, rather than positively, associated with fern species richness (Table 1). Our analysis suggests that the results of the climate-based analyses reported in Suissa et al. (2021) are likely biased to a large degree.

We conclude that fern species lists derived from GBIF are generally very incomplete across a wide range of spatial scales (at least <300,000 km²), and should not be used in studies that require data derived from complete or nearly complete species lists. This conclusion likely applies to data with GBIF for all taxonomic groups of organisms.

### Author contributions

H.Q. designed research, analyzed data, and wrote the paper; J.Z. prepared data; M.J. generated maps; all authors participated in revising the paper.

### Declaration of competing interest

The authors declare no conflict of interest.

### Acknowledgements

We thank two reviewers for their helpful comments.

### Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.pld.2021.10.001.

### References

Adersen, H., 1995. Research on islands: classic, recent, and prospective approaches. In: Vitousek, P.M., Loope, L.L., Adersen, H. (Eds.). Islands: Biological Diversity and Ecosystem Function. Springer-Verlag, Berlin, pp. 7–21.

Beck, J., Ballesteros-Mejia, L., Nagel, P., et al., 2013. Online solutions and the ‘Wallacean shortfall’: what does GBIF contribute to our knowledge of species’ ranges? Divers. Distrib. 19, 1043–1050.

Brunnitt, R.K., 2001. World Geographical Scheme for Recording Plant Distributions, 2 edn. Hunt Institute for Botanical Documentation, Carnegie Mellon University, Pittsburgh.

Feng, G., Mao, L., Sandel, B., et al., 2016. High plant endemism in China is partially linked to reduced glacial-interglacial climate change. J. Biogeogr. 43, 145–154.

Hassler, M., 2004–2021. World Ferns: Synonymic Checklist and Distribution of Ferns and lycophytes of the World, Version 12.3. www.worldplants.de/ferns/.

Kissling, W.D., Carl, G., 2008. Spatial autocorrelation and the selection of simultaneous autoregressive models. Global Ecol. Biogeogr. 17, 59–71.

Kooyman, R., Rossetto, M., Allen, C., et al., 2012. Australian tropical and subtropical rain forest community assembly: phylogenetic functional biogeography, and environmental gradients. Biotropica 44, 668–679.

Kreft, H., Jetz, W., Mutke, J., et al., 2010. Contrasting environmental and regional effects on global pteridophyte and seed plant diversity. Ecology 91, 408–419.

Mabberley, D.J., 2008. Mabberley’s Plant-Book: A Portable Dictionary of Plants, Their Classifications, and Uses, third ed. Cambridge University Press, Cambridge.

Nagalingum, N.S., Knerr, N., Laffan, S., et al., 2015. Continental scale patterns and predictors of fern richness and phylogenetic diversity. Front. Genet. 6, 132.

Qian, H., 2009. Beta diversity in relation to dispersal ability for vascular plants in China. J. Biogeogr. 36, 1050–1058.

Qian, H., Jin, Y., Ricklefs, R.E., 2017. Phylogenetic diversity anomaly in angiosperms ranges? Divers. Distrib. 19, 1043–1050.

Khine, P.K., Kluge, J., Kessler, M., et al., 2019. Latitude-independent, continent-wide consistency in climate–richness relationships in Asian ferns and lycophytes. J. Biogeogr. 46, 581–591.

Kissling, W.D., Carl, G., 2008. Spatial autocorrelation and the selection of simultaneous autoregressive models. Global Ecol. Biogeogr. 17, 59–71.

Kooyman, R., Rossetto, M., Allen, C., et al., 2012. Australian tropical and subtropical rain forest community assembly: phylogenetic functional biogeography, and environmental gradients. Biotropica 44, 668–679.

Kreft, H., Jetz, W., Mutke, J., et al., 2010. Contrasting environmental and regional effects on global pteridophyte and seed plant diversity. Ecology 91, 408–419.

Mabberley, D.J., 2008. Mabberley’s Plant-Book: A Portable Dictionary of Plants, Their Classifications, and Uses, third ed. Cambridge University Press, Cambridge.

Nagalingum, N.S., Knerr, N., Laffan, S., et al., 2015. Continental scale patterns and predictors of fern richness and phylogenetic diversity. Front. Genet. 6, 132.

Qian, H., 2009. Beta diversity in relation to dispersal ability for vascular plants in China. J. Biogeogr. 36, 1050–1058.

Qian, H., Jin, Y., Ricklefs, R.E., 2017. Phylogenetic diversity anomaly in angiosperms ranges? Divers. Distrib. 19, 1043–1050.

Khine, P.K., Kluge, J., Kessler, M., et al., 2019. Latitude-independent, continent-wide consistency in climate–richness relationships in Asian ferns and lycophytes. J. Biogeogr. 46, 581–591.

Kissling, W.D., Carl, G., 2008. Spatial autocorrelation and the selection of simultaneous autoregressive models. Global Ecol. Biogeogr. 17, 59–71.

Kooyman, R., Rossetto, M., Allen, C., et al., 2012. Australian tropical and subtropical rain forest community assembly: phylogenetic functional biogeography, and environmental gradients. Biotropica 44, 668–679.

Kreft, H., Jetz, W., Mutke, J., et al., 2010. Contrasting environmental and regional effects on global pteridophyte and seed plant diversity. Ecology 91, 408–419.

Mabberley, D.J., 2008. Mabberley’s Plant-Book: A Portable Dictionary of Plants, Their Classifications, and Uses, third ed. Cambridge University Press, Cambridge.

Nagalingum, N.S., Knerr, N., Laffan, S., et al., 2015. Continental scale patterns and predictors of fern richness and phylogenetic diversity. Front. Genet. 6, 132.
Qian, H., Zhou, Y., Zhang, J., et al., 2021b. A synthesis of botanical informatics for vascular plants in Africa. Ecol. Inf. 64, 101382.

Stohlgren, T.J., Barnett, D.T., Kartesz, J.T., 2003. The rich get richer: patterns of plant invasions in the United States. Front. Ecol. Environ. 1, 11–14.

Suisse, J.S., Sundue, M.A., Testo, W.L., 2021. Mountains, climate and niche heterogeneity explain global patterns of fern diversity. J. Biogeogr. 48, 1296–1308.

Weigand, A., Abrahamczyk, S., Aubin, I., et al., 2020. Global fern and lycophyte richness explained: how regional and local factors shape plot richness. J. Biogeogr. 47, 59–71.

Wegelt, P., Kissling, W.D., Kisel, Y., et al., 2015. Global patterns and drivers of phylogenetic structure in island floras. Sci. Rep. 5, 12213.

Weiser, M.D., Swenson, N.G., Enquist, B.J., et al., 2018. Taxonomic decomposition of the latitudinal gradient in species diversity of North American floras. J. Biogeogr. 45, 418–428.

Wolf, P.G., Schneider, H., Ranker, T.A., 2001. Geographic distribution of homosporous ferns: does dispersal obscure evidence of vicariance? J. Biogeogr. 28, 263–270.

Yesson, C., Brewer, P.W., Sutton, T., et al., 2007. How global is the global biodiversity information facility? PLoS One 2, e1124.

Zhang, J., Qian, H., Girardello, M., et al., 2018. Trophic interactions among vertebrate guilds and plants shape global patterns in species diversity. Proc. Roy. Soc. Lond. B 285, 20180949.