The use of big data and data science techniques in environmental science

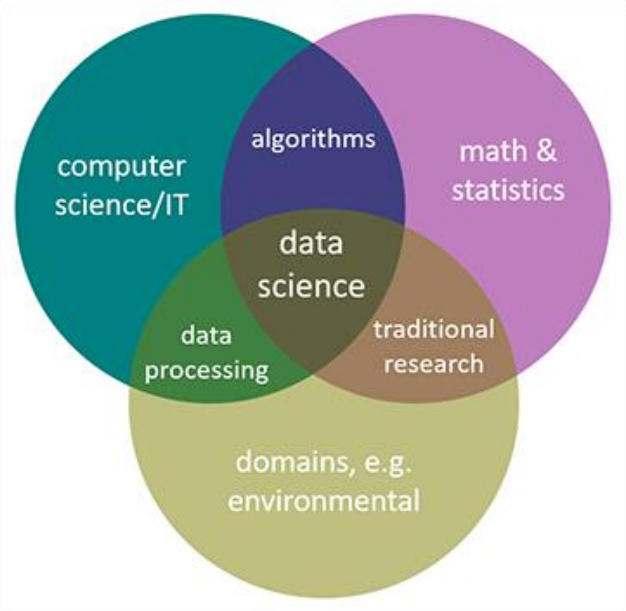

Figure 1: Different aspects in the study of environmental data science

Table of Contents

Word Count: 3099

Abstract

With the increasing volume of ecological data, the field of environmental science is rapidly advancing as, ecological modelling, image classification and the use of remote sensing have all influenced the field of environmental science. Ecology studies now use large online datasets to help to test hypothesis in ecology as well as monitor and track trends in ecological and evolutionary processes. Hydrology and climatology studies have made use of long-term data sets on rainfall and temperatures to help to describe the extent of droughts, flooding and extreme temperatures and linking these extremes to anthropogenic climate changes. Anthropogenic pressures on the environmental have caused rapid environmental changes, with data science techniques helping to monitor these rapid changes. The field of environmental science faces a variety of challenges, which can be addressed with the use of data science and data driven techniques, though challenges still exist, particularly in the modelling of complex systems.

1. Introduction

The use of big data has risen exponentially in the field of environmental science in a variety of dimensions including size, resolution, and complexity (Gibert, 2018). This data is collected from a range of sources including remote sensing; earth monitoring systems; field campaigns; historic records; model output and data mining from social media platforms (Blair, 2019). This expansion in data is helping to advancing the field in a variety of topics in environmental science, including ecological sciences, hydrological sciences, and the human impacts on earths systems - with a new journal focusing on environmental data science publishing it's first papers in 2022 (Monteleoni, 2022). Moreover, tools such as machine learning are increasingly being used in environmental studies, with the term rising exponentially in the literature, especially in the hydrosphere (Zhong, 2021). While the use of data science has advanced the environmental field a range of ethical and social issues exist. For example, biodiversity datasets, such as those collected by GBIF (Global biodiversity information facility) (Meyer, 2015), have a higher density of data in spaces with higher urbanisation and economically developed countries with high internet access: researchers also have more financial incentive to digitise records in high income countries leading to biases and gaps in biodiversity data. Sensors and monitoring systems can also often be faulty, especially in extreme climates. Gathering and storing environmental and ecological data may also have a variety of negative environmental effects, leading to increased pollution and resource depletion, such as the use of energy, clean water, and rare materials to operate data centres (McGovern, 2022).

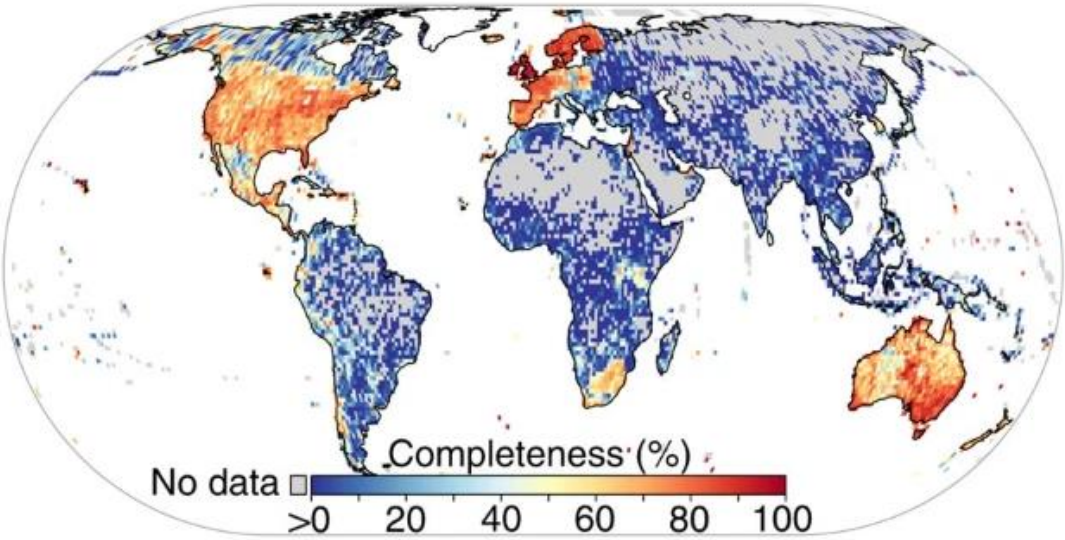

Figure 2: Percent of data digitised to online records on GBIF (Global biodiversity information facility)

Three areas of data science - modelling, image classification and the use of remote sensing have all been utilised to enhance environmental research (Perry, 2022), as well as the organisation of data into large databases. A range of software tools have been developed to optimise data management and visualisation. Software for GIS(geographical information systems) has expanded with ESRI and QGIS3 the most popular tools for modelling geospatial data (GISGeography, 2023), while R programming language is now the most popular programming language for ecologists (Lai, 2019), as well as hydrologists (Slater, 2019), with the package *data.table* (Dowle, 2023) allowing for faster processing of large datasets.

2. Ecology and biodiversity

Being able to measure and quantify biodiversity is important so monitoring can be carried out to track the growing threats of biodiversity loss due to anthropogenic pressures. However, biodiversity has several dimensions, including species richness, phylogenetic, functional traits and number of interactions in an ecosystem (Pollock, 2020). Hence, reducing biodiversity to a single metric only gives a limited view. This has caused a range of issues, for example the vulnerability of monoculture trees and mangrove restoration to the spread of a range of diseases (Hai, 2020). There are a wide range of gaps in data collection, both spatially and taxonomically (Troudet, 2017)(Fig 3.), nonetheless a range of models have been created using the current data to gain insight into biodiversity patterns and trends.

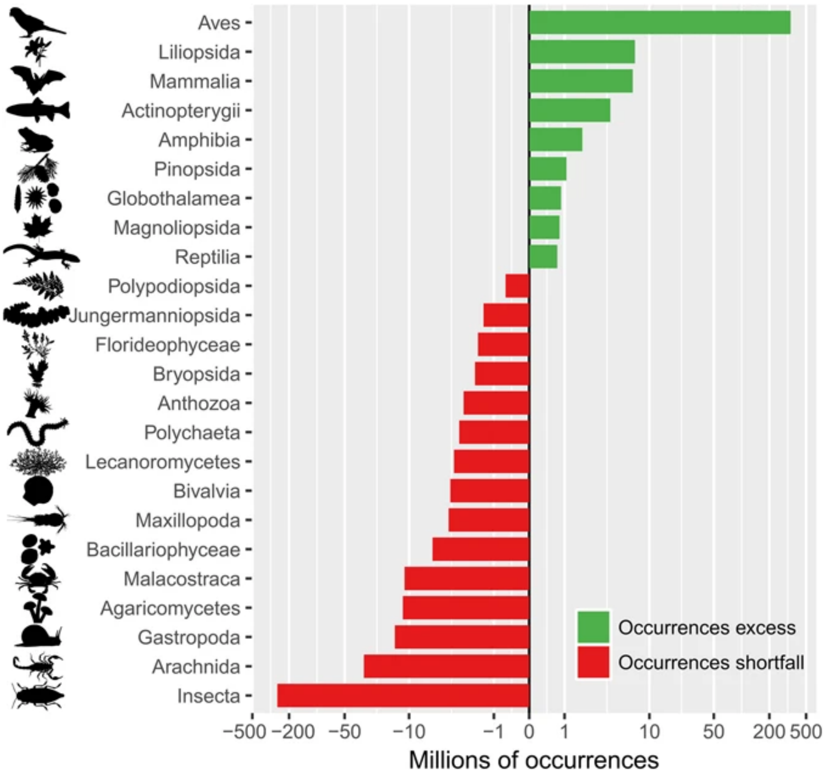

Figure 3: Gaps in taxonomic biodiversity data

2.1. Ecological modelling and forecasting

Advanced machine learning techniques were used to map global terrestrial diversity of vascular plant species (Cai, 2022), using an online catalogue of plants as the baseline data. The study helped to show that higher variation in plant biodiversity was caused by climate and water availability, relative to habitat heterogeneity. Both species richness and functional diversity were mapped and exposed the non-uniform diversity of terrestrial vascular plants, identifying specific regions known as "hotspots" with a significantly higher amount of biodiversity. The study was able to give a finer level of detail, particularly in regions with steep elevation gradients relative to previous studies, though at extreme points in space, where few species were present, smaller scale models had a stronger performance.

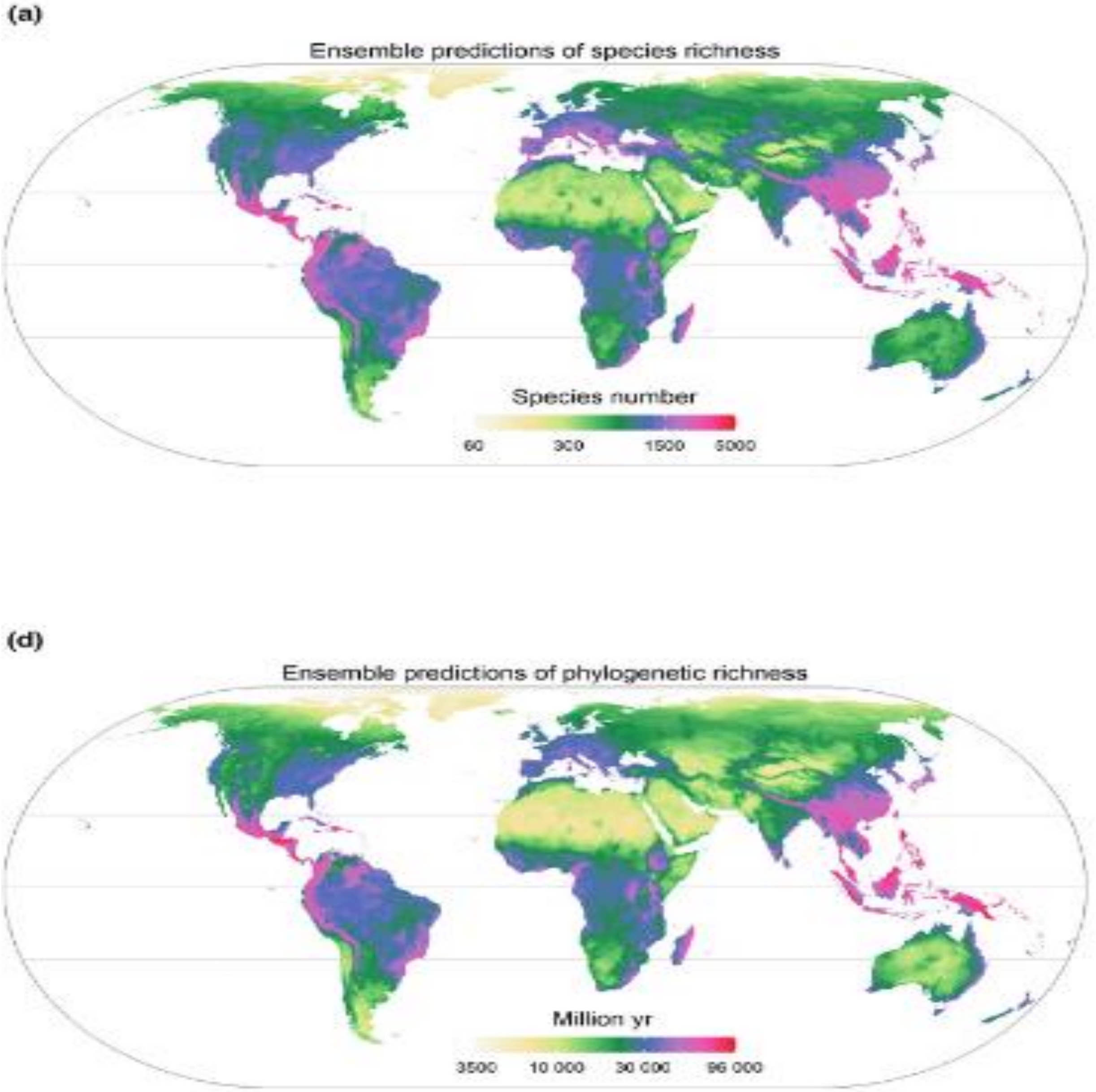
(a)
Ensemble predictions of species richness
Species number
60
300
1500
6000
(d)
Ensemble predictions of phylogenetic richness
Million yr
3500
10 000
30 000
96 000

Figure 4:Global spread of species and phylogenetic plant diversity

Databases are being designed to record a range of ecological phenomena, for example how plants respond when exposed to different mycorrhizal fungi (Chaudhary, 2016). The database is biased towards plants which are important for agriculture and fungi taxa which have been commercially marketed, nonetheless the database may be helpful to refer to in future studies of the influence of fungi on plant productivity. Similarly, a database of patterns of global fine root morphology (Iversen, 2017) has been created to help to investigate the effect of fine root morphology on plant development and growth characteristics.

Eco acoustics is a relatively new field helping to advance ecological modelling and conservation, with new, cheaper sensors helping to capture data in a variety of remote locations (Sheng, 2019). Being able to interpret differences between human noise, wildlife noise, and weather phenomena has been achieved (Quinn, 2022)- using deep learning models, suggesting acoustic classification of broad soundscapes is possible. Eco acoustics can also be used below soil, to monitor soil biodiversity (Robinson, 2023) and in freshwater systems such as ponds to monitor aquatic biodiversity (Linke, 2018).

Monitoring and tracking mammals by collecting in the field data is time-consuming, labour-intensive, and expensive. Therefore a range of stationary and mobile sensors(Fig 5.) could be utilised with machine learning models in the future to help ecologist monitor ecological systems, a range of corporate and research organisation have developed models in animal ecology, which can be utilised in the field of conservation, though some biological processes have not yet been integrated into the machine learning models (Tuia, 2022). One such biological process is biotic interactions between species. Species distribution models have been designed to predict the potential range of a species habitat which make use of abiotic and biotic data, nonetheless the integration of biotic data remains a challenge (Zimmermann, 2010).

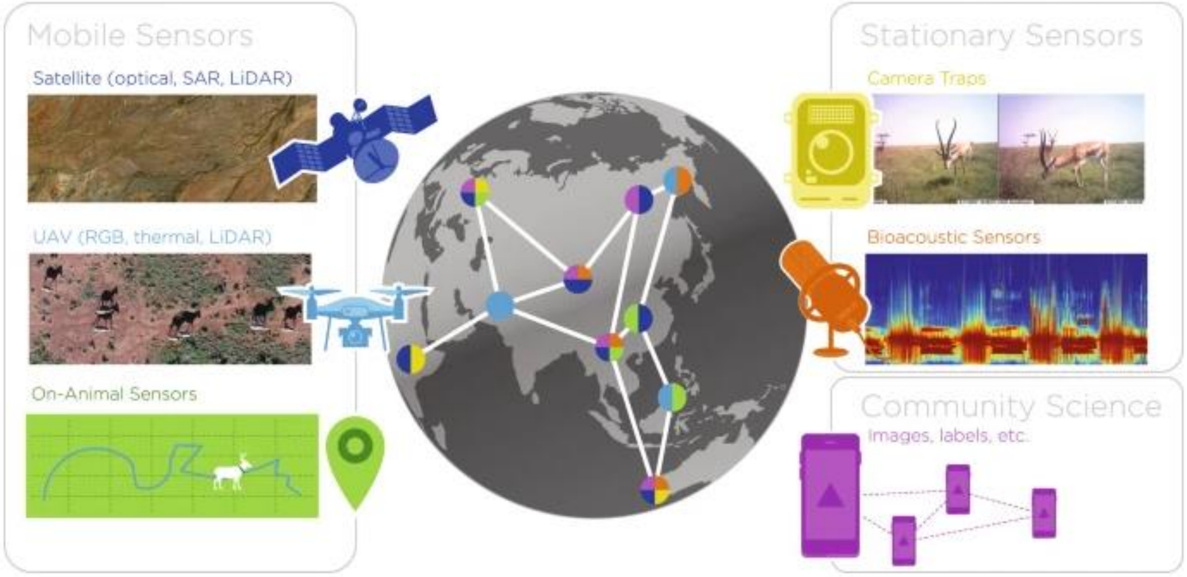

Figure 5: Mobile and Stationary sensors to collect animal ecological data.

2.2. Image detection and classification

Being able to classify species using image data is important for ecologists to monitor the evolution of species within an ecological community. Image classification techniques can be applied across a variety of species to monitor ecological and evolutionary processes.
Machine learning has been used on image data collected by ecologists to classify species of plankton (Orenstein, 2022), using morphological and physiological features of the planktonic species. The technique could also be applied to a range of marine and freshwater organisms, using image and video data, as well as expand knowledge on planktonic species functional diversity and species traits (Ryabov, 2022).

The use of camera traps to sample ecological habitats is well known, nonetheless classification of species of arthropods is generally not possible. However, the use of convolution neural networks has been used to classify ground bark beetles (Coleoptera: Carabidae), using images from the natural history museum as the ground truth (Robinson, 2013). Use of image classification can also be used to identify plant stress and plant diseases, based on leaf morphology and colour, hence this is important in a range of areas from agriculture to conservation (Lowe, 2017), as well as helping select plant species which can deal with abiotic and biotic stresses (Singh, 2021).

2.3. Remote sensing

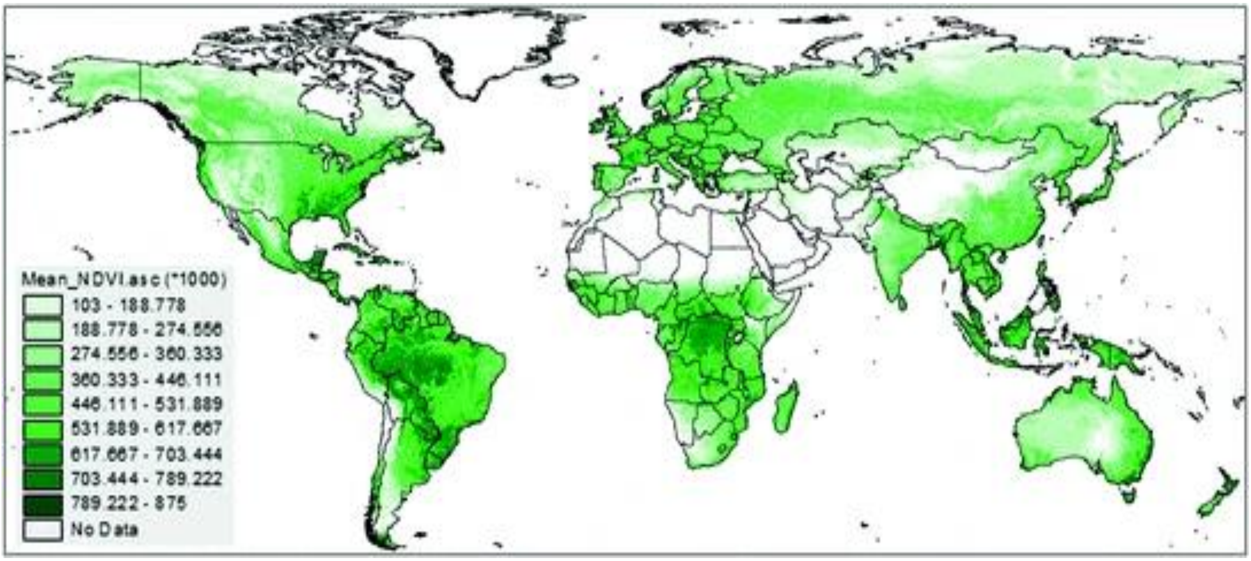

Figure 6: Mean NDVI values period 1982-2006

Remote sensing has transformed the field of ecology, due to the cheap means of accessing data. Many aspects of remote sensing have improved over the past decades, such as spatial resolution(<10m), the range of spectral data observed (Hyperspectral), thermal remote sensing and LIDAR (Wang, 2010). Moreover, the use of Unmanned aerial vehicles (UAVs) also allows for the investigation of a range of remote or inaccessible habitats at high resolution.

LIDAR is an important tool to measure sub canopy vegetation in forests (Jarron, 2020), which cannot be measured accurately with legacy techniques. The composition of sub canopy vegetation is important as it is a key driver for forest succession dynamics and can help to quantity a range of ecosystem services provided by forests.

Remote sensing can capture a variety of indices using different wavelengths of light. One of the most popular indices is the normalised differential vegetation index (NDVI), which can be calculated from satellite data from Sentinel 2a or Landsat. The number of papers discussion NDVI rose from 800 in the 1990s to 12,000 in the 2010, highlighting the importance of this index in environmental research (Huang, 2021). NDVI was used to show disturbances in vegetation, due to changing abiotic and biotic factors, such as fire, drought, disease, though NDVI is not a good measure for sub canopy vegetation. NDVI is also hypothesised as a key metric to measure spectral diversity, which according to the spectral variability hypothesis, is an effective measure of plant diversity, though research suggests, other factors such as productivity, climate, and historical processes also influence plant diversity, hence NDVI alone is not a good sole proxy for plant biodiversity (Perrone, 2023).

3. Climatology and hydrology

Extreme events in climate and the hydrological cycle are occurring more frequently due to climate change and changing aerosol loading, consequently monitoring changes are important for a range of human systems such as weather forecasting and early warning of extreme weather events such as flooding, droughts, hurricanes, and wildfires.

3.1. Modelling and forecasting

Drought forecasting is an important tool for scientists and policy makers, due to the high costs of drought to human societies. Drought can be defined in several ways, including agricultural drought - due to low soil moisture or hydrological drought - due to low river and lake levels. A range of models exist to predict droughts occurrences and severity. Over a 250 year period from 1766-2015, the mesoscale hydrological model was used to show the changing nature of drought in Europe (Moravec, 2019), the results showed that in Mediterranean regions, soil moisture droughts were increasing in frequency and extremity, while in central Europe - hydrological droughts were declining in spatial extent. Increasing temperatures and lack of rainfall in the summer months were leading to greater evapotranspiration and were characteristic of recent summer droughts in Europe in 2003 and 2015. A web application was also produced to visualise the results.

Similarly, models have been constructed to analyse precipitation over time, with higher resolution visualisation now possible, such as the Multi-Source Weighted-Ensemble Precipitation v2.0. Changes in precipitation over land were investigated between the period 1979-2016, analysing the intensification of the hydrological cycle hypothesis, stating that drier places get drier and wetter places get wetter due to the strong nonlinear dependence of water vapour pressure (Martinkova, 2020). The results showed, an increase in total precipitation, number of wet days and heavy events (Markonis, 2019) across all continents except South America where the number of wet days declined. Increases in precipitation were particularly large over the inter tropical convergence zone, while decreased precipitation was observed over the sub tropics.

Groundwater is the largest source of freshwater on earth, as well as supporting freshwater ecosystems in times of drought. A model called MODFLOW was used to estimate worldwide groundwater levels. Validation was completed from observed groundwater discharge. The model appreciated well in sediment basins, though overestimated water levels on steeper terrain (de Graaf, 2015).

Climate and weather modelling have been strongly influenced by data driven techniques in the past decades. Weather modelling on a weekly or monthly timescale still has various issues due to high uncertainty in physical processes in cloud systems (Chantry, 2021). Drivers of climate change models also have many uncertainties, especially due to aerosol optical and microphysical properties (Li, 2022) and there interactions with clouds, nonetheless machine learning techniques are helping to reduce uncertainty (Chen, 2022). A range of data sources have also been utilised to model past climates including tree rings (Evans, 2006), ice cores and sedimentary rock.

3.2. Image detection, classification, and use of remote sensing

Lakes across the Tibetan plateau are hard to access and get in-situ data, therefore the use of remote sensing is vital to observe the changing structure and size of Tibetan lakes due to climate change, as well as the potential for monitoring glacial lake outbursts floods, which cause significant damages to mountain communities (Song, 2014).

Water quality of inland waters can be detected using remote sensing data. In the past data would have to be collected on site, however, with remote sensing data, a training dataset was made to test for the levels of chlorophyl A present in inland lakes (Pulliainen, 2001). The use of hyperspectral remote sensing data, with higher spatial, spectral, and temporal resolution can also be used with machine learning to monitor water quality for lakes and rivers (Sun, 2022).

Predicting root soil moisture(>5cm), is challenging due the non linear relationship between surface soil moisture and root soil moisture, while predicting surface soil moisture using remote sensing techniques is well researched (Li, 2021), predicting deeper levels was improved with data driven techniques (Yinglan, 2022), using remote sensed data to work out the normalised vegetation differential index(NDVI) as the predictive variable in the model. The use of remotely sensed soil moisture was also a key variable to predict the extent of wildfires across the Iberian Peninsula (Chaparro, 2016). Research using remote sensing and wildfires has grown significantly in the last decade (Santos, 2021), with remotely sensed data providing a plethora of variables which can be used with data driven techniques to model forest fires (Sayad, 2019).

4. Global environmental change

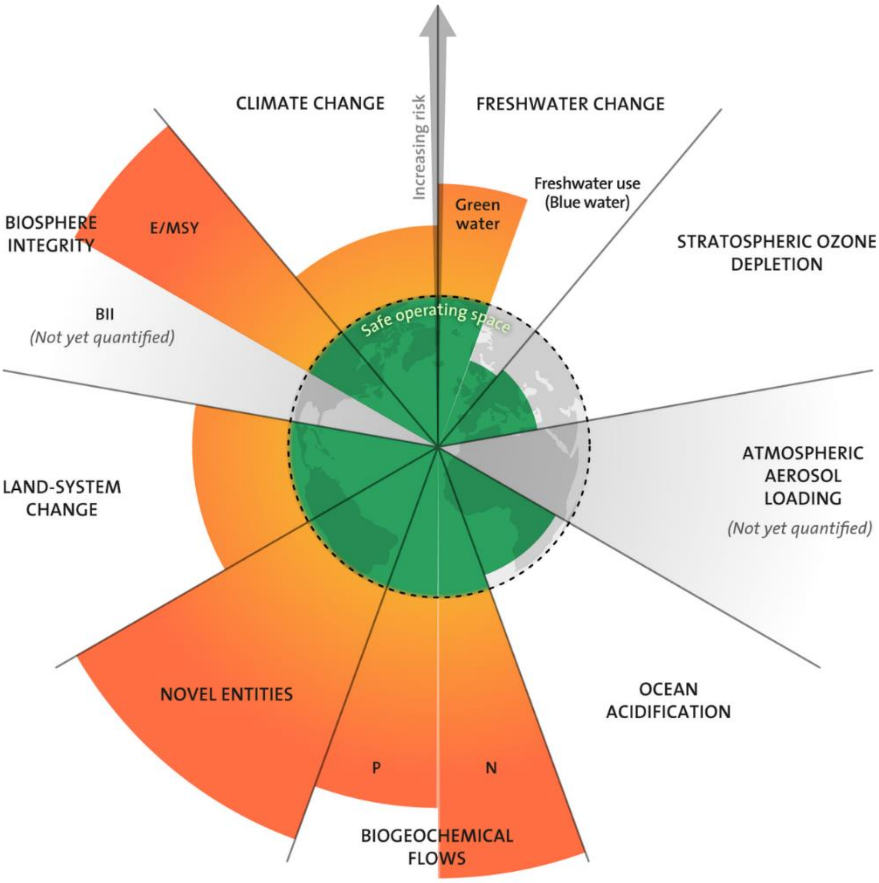

Figure 7: Planetary boundaries framework

The last 12000 years have been characterised by relatively stable weather patterns, termed the Holocene, which have allowed for the rapid development of human societies, with the neolithic revolution and the industrial revolution having the widest impacts on human societies. However, these developments have caused rapid changes to earth's climate and environmental systems on a global scale, with the last 200 years termed the Anthropocene due to the immense changes taking place due to human activities. The planetary boundaries framework has been constructed to highlight earth systems at risk of tipping into unstable systems, including climate, biogeochemical cycling, and land use systems (Fig 7.).

Wetlands which were once the most dominant ecosystem on land have been cleared to make way for agriculture, urban spaces, and ease of river navigation. Land use and land change has taken place rapidly, as human societies have expanded. Urban expansion or “urban sprawl” has occurred as humans have moved from villages to cities. Remote sensing data can be combined with social media data to track urban sprawl (Shao, 2021) leading to a range of ecosystem disservices. A random forest algorithm was used with 4 classes: built up, vegetation, agriculture, and water. the random forest model was optimal due to requiring fewer parameters, minimal manual intervention, and yielded a high classification accuracy, and could also manage higher-dimensional data and obtain classification results rapidly.

Remote sensing has also been used to map agricultural sprawl, due to changing diets and increasing population. Our world in data shows that 4.9 billion hectors of land are now used for animals agriculture and crops, with exponential rises in land use since the industrial revolution (Ritchie, 2013), NDVI indices were measured using a decision tree classifier from 1985-2018 in MATOPIBA region, in eastern Brazil, showing reduced natural vegetation, soil erosion and land degradation, due to unsustainable anthropogenic activities including fire clearance and the removal of forest ecosystems (Vieira, 2021), Agricultural activity in desert environments has also put pressure on groundwater, leading to lack of buoyancy at the land surface and increasing number of sinkholes (Youssef, 2019).

Ozone depletion was a well discussed topic in the 1980s, because of CFCs (Chlorofluorocarbons) damaging the ozone layer, since the Montreal protocol banning CFCs, the ozone hole has improved, nonetheless a variety of chemical substance influence the ozone layer. One model tried to predict the extent of the ozone present in the upper atmosphere from data published in our world in data and compared a range of machine learning model to predict the ozone concentration. From the models the support vector machine achieved the lowest mean square error (Dong, 2023), though other factors may also influence the ozone layer such as forest fires and nitrous dioxide emissions, not account for in the model. Knowledge about the ozone concentration profile between 1979-2020 was also modelled, using a chemical transport model and random forest ensemble learning, as there is no long-term data on ozone profile from a single satellite (Dhomse, 2021).

Humans have also changed a variety of environmental microbiomes. At present it is thought that only a small fraction of earth's microbe diversity has been recorded. Microbiome studies are now utilising a range of deep learning techniques in areas such as microbiome engineering to improve crop yields or improve human health (Hernández Medina, 2022).

5. Conclusions

The use of data science techniques in environmental science has developed and enhanced the study of environmental sciences, with image classification, modelling and remote sensing all influencing environmental sciences. In the future the spatial, temporal, and spectral resolution of images from new satellites and UAVs such as drones will help to increase the quality of data sources and fundamental questions such as the spectral hypothesis of biodiversity will be easier to investigate and verify. Data volume has also been influenced by citizen science projects in the field of ecology, helping to follow a range of trends such as insect loss in the Anthropocene (Sky News, 2022). Early warning systems to alert citizens to environmental disasters, such as floods, droughts and wildfires may also improve and help policy makers to make informed decisions on constructing resilient infrastructure to mitigate environmental risks.

A range of challenges exist in the field of environmental data science (Blair, 2019), such as the need for cultural shift in the use open data and to discover new modes of cross collaboration between disciplines, as well to being able to explain complex systems which occur frequently in the domain of environmental science.

6. References

1) Blair, G. e. a., 2019. Data Science of the Natural Environment: A Research Roadmap. *Frontiers in Environmental Science,* Volume 7.

2) Cai, L. e. a., 2022. Global models and predictions of plant diversity based on advanced machine learning techniques. *New Phytologist,* Volume 237, pp. 1432-1445.

3) Chantry, M. a., 2021. Opportunities and challenges for machine learning in weather and climate modelling: hard, medium and soft AI. *Phil.Trans.R.So,* Volume 379.

4) Chaparro, D. e. a., 2016. Predicting the Extent of Wildfires Using Remotely Sensed Soil Moisture and Temperature Trends. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* 9(6), pp. 2818-2829.

5) Chaudhary, V. e. a., 2016. MycoDB, a global database of plant response to mycorrhizal fungi. *Sci Data,* Volume 3.

6) Chen, Y. e. a., 2022. Machine learning reveals climate forcing from aerosols is dominated by increased cloud cover. *Nat. Geosci,* Volume 15, pp. 609-614.

7) de Graaf, I. e. a., 2015. A high-resolution global-scale groundwater model. *Hydrol. Earth Syst. Sci,* Volume 19, p. 823–837.

8) Dhomse, S. e. a., 2021. ML-TOMCAT: machine-learning-based satellite-corrected global stratospheric ozone profile data set from a chemical transport model. *Earth System Science Data,* Volume 13, p. 5711–5729.

9) Dong, Y., 2023. Regression-based Analysis of Ozone Layer via Machine Learning Models. *Highlights in Science, Engineering and Technology,* Volume 39, p. 1356–1363.

10) Dowle, M., 2023. *data.table.* [Online]
Available at: https://www.rdocumentation.org/packages/data.table/versions/1.14.8
[Accessed 29 04 2023].

11) Emery, N. a., 2021. Data Science in Undergraduate Life Science Education: A Need for Instructor Skills Training. *BioScience,* 71(12), p. 1274–1287.

12) Evans, M. N. e. a., 2006. A forward modeling approach to paleoclimatic interpretation of tree-ring data. *J. Geophys. Res.,* 111(G3).

13) Gibert, K. e. a., 2018. Environmental Data Science. *Environmental Modelling & Software,* Volume 106, pp. 4-12.

14) GISGeography, 2023. *30 Best GIS Software Applications [Rankings].* [Online]
Available at: https://gisgeography.com/best-gis-software/
[Accessed 01 05 2023].

15) Hai, N. e. a., 2020. Towards a more robust approach for the restoration of mangroves in Vietnam. *Annals of Forest Science,* Volume 77.

16) Hernández Medina, R. K. S. N. K. e. a., 2022. Machine learning and deep learning applications in microbiome research. *ISME COMMUN,* 2(98).

17) Huang, S. e. a., 2021. A commentary review on the use of normalized difference vegetation index (NDVI) in the era of popular remote sensing. *Forestry Research,* Volume 32, pp. 1-6.

18) Iversen, C. e. a., 2017. A global Fine-Root Ecology Database to address below-ground challenges in plant ecology. *New Phytol,* Volume 215, pp. 15-26.

19) Jarron, L. e. a., 2020. Detection of sub-canopy forest structure using airborne LiDAR. *Remote Sensing of Environment,* Volume 244.

20) Lai, J. e. a., 2019. Evaluating the popularity of R in ecology. *Ecosphere,* Volume 10.

21) Li, J. e. a., 2022. Scattering and absorbing aerosols in the climate system. *Nat Rev Earth Environ,* Volume 3, pp. 363-379.

22) Linke, S., 2018. Freshwater ecoacoustics as a tool for continuous ecosystem monitoring. *Front Ecol Environ,* 16(4).

23) Li, Z., 2021. Soil moisture retrieval from remote sensing measurements: Current knowledge and directions for the future. *Earth-Science Reviews,* Volume 218.

24) Lowe, A. e. a., 2017. Hyperspectral image analysis techniques for the detection and classification of the early onset of plant disease and stress. *Plant Methods,* Volume 13.

25) Markonis, Y. e. a., 2019. Assessment of Water Cycle Intensification Over Land using a Multisource Global Gridded Precipitation DataSet. *JGR Atmostpheres,* 124(21), pp. 11175-11187.

26) Martinkova, M. K. J., 2020. Overview of Observed Clausius-Clapeyron Scaling of Extreme Precipitation in Midlatitudes. *Atmosphere,* 11(8), p. 786.

27) McGovern, A. e. a., 2022. Why we need to focus on developing ethical, responsible, and trustworthy artificial intelligence approaches for environmental science. *Environmental Data Science,* Volume 1, p. e6.

28) Meyer, C. e. a., 2015. Global priorities for an effective information basis of biodiversity distributions. *Nature Communications,* Volume 6.

29) Monteleoni, C., 2022. *Environmental Data Science.* [Online]
Available at: https://www.cambridge.org/core/journals/environmental-data-science
[Accessed 28 04 2023].

30) Moravec, V. a., 2019. A 250-Year European Drought Inventory Derived From Ensemble Hydrologic Modeling. *geophysucal research letters,* 46(11), pp. 5909-5917.

31) Orenstein, A. e. a., 2022. Machine learning techniques to characterize functional traits of plankton from image data. *Limnol Oceanogr,* Volume 67, pp. 1647-1669.

32) Perrone, M. e. a., 2023. The relationship between spectral and plant diversity: Disentangling the influence of metrics and habitat types at the landscape scale. *Remote Sensing of Environment,* Volume 293.

33) Perry, G. e. a., 2022. An Outlook for Deep Learning in Ecosystem Science. *Ecosystems,* Volume 25, p. 1700–1718.

34) Pollock, L. e. a., 2020. Protecting Biodiversity (in All Its Complexity): New Models and Methods. *trends in Ecology and Evolution,* 35(12), pp. 1119-1128.

35) Pulliainen, J. e. a., 2001. A semi-operative approach to lake water quality retrieval from remote sensing data. *Science of The Total Environment,* 268(1-3), pp. 79-93.

36) Quinn, C. e. a., 2022. Soundscape classification with convolutional neural networks reveals temporal and geographic patterns in ecoacoustic data. *Ecological Indicators,* Volume 138.

37) Ritchie, H. R. M., 2013. *Land Use.* [Online]
Available at: https://ourworldindata.org/land-use
[Accessed 29 04 2023].

38) Robinson, J. e. a., 2013. Interspecific synchrony of seabird population growth rate and breeding success. *Ecology and Evolution,* 3(7), p. 2013– 2019.

39) Robinson, J. e. a., 2023. The sound of restored soil: Measuring soil biodiversity in a forest restoration chronosequence with ecoacoustics. *bioRxiv.*

40) Ryabov, A. e. a., 2022. Estimation of functional diversity and species traits from ecological monitoring data. *PNAS ecology,* Volume 43, p. 119.

41) Santos, S. e. a., 2021. Research on Wildfires and Remote Sensing in the Last Three Decades: A Bibliometric Analysis. *Forests,* 12(5), p. 604.

42) Sayad, O. e. a., 2019. Predictive modeling of wildfires: A new dataset and machine learning approach. *Fire Safety Journal,* Volume 104, pp. 130-146.

43) Shao, Z. e. a., 2021. Urban sprawl and its impact on sustainable urban development: a combination of remote sensing and social media data. *Geo-spatial Information Science,* 24(2), pp. 241-255.

44) Sheng, Z. e. a., 2019. Wireless acoustic sensor networks and edge computing for rapid acoustic monitoring. *IEEE/CAA Journal of Automatica Sinica,* 6(1), pp. 64-74.

45) Singh, A. e. a., 2021. Challenges and Opportunities in Machine-Augmented Plant Stress Phenotyping. *Trends in plant science,* 26(1), pp. 53-69.

46) Sky News, 2022. *Car number plate 'splatometer' survey shows 'terrifying decline' in number of flying insects.* [Online]
Available at: https://news.sky.com/story/car-number-plate-squashed-bug-survey-shows-terrifying-decline-in-number-of-flying-insects-12605662
[Accessed 1 5 2023].

47) Slater, L. e. a., 2019. Using R in hydrology: a review of recent developments and future directions. *Hydrology and Earth System Sciences,* 23(7), p. 2939–2963.

48) Song, C. e. a., 2014. Remote sensing of alpine lake water environment changes on the Tibetan Plateau and surroundings: A review. *ISPRS Journal of Photogrammetry and Remote Sensing,* Volume 92, pp. 26-37.

49) Sun, X. e. a., 2022. Monitoring water quality using proximal remote sensing technology. *Science of The Total Environment,* Volume 803.

50) Troudet, J. e. a., 2017. Taxonomic bias in biodiversity data and societal preferences. *Sci Rep,* Volume 7, p. 9132.

51) Tuia, D. e. a., 2022. Perspectives in machine learning for wildlife conservation. *Nat Commun,* Volume 13, p. 792.

52) Vieira, R. e. a., 2021. Land degradation mapping in the MATOPIBA region (Brazil) using remote sensing data and decision-tree analysis. *Science of The Total Environment,* Volume 782.

53) Wang, K. e. a., 2010. Remote Sensing of Ecology, Biodiversity and Conservation: A Review from the Perspective of Remote Sensing Specialists. *Sensors,* 10(11), pp. 9647-9667.

54) Yinglan, A. e. a., 2022. Root-zone soil moisture estimation based on remote sensing data and deep learning. *Environmental Research,* Volume 212.

55) Youssef, A. e. a., 2019. Agriculture Sprawl Assessment Using Multi-Temporal Remote Sensing Images and Its Environmental Impact. *Sustainability,* 11(15), p. 4177.

56) Zhong, S. e. a., 2021. Machine Learning: New Ideas and Tools in Environmental Science and Engineering. *Environ. Sci. Technol.,* 55(19), p. 12741–12754.

57) Zimmermann, E. e. a., 2010. New trends in species distribution modelling. *Ecography,* 33(6), pp. 985-989.

7. Image sources

I. Figure 1) Introduction to Environmental Data Science, Jerry Davies, https://bookdown.org/igisc/EnvDataSci/, accessed 01/05.2023

II. Figure 2) Meyer, C., Kreft, H., Guralnick, R. *et al.* Global priorities for an effective information basis of biodiversity distributions. *Nat Commun* **6**, 8221 (2015). https://doi.org/10.1038/ncomms9221

III. Figure 3) Troudet, J., Grandcolas, P., Blin, A. *et al.* Taxonomic bias in biodiversity data and societal preferences. *Sci Rep* **7**, 9132 (2017). https://doi.org/10.1038/s41598-017-09084-6

IV. Figure 4) Cai, L., Kreft, H., Taylor, A. et al. (2023), Global models and predictions of plant diversity based on advanced machine learning techniques. New Phytol, 237: 1432-1445. https://doi.org/10.1111/nph.18533

V. Figure 5) Tuia, D., Kellenberger, B., Beery, S. *et al.* Perspectives in machine learning for wildlife conservation. *Nat Commun* **13**, 792 (2022). https://doi.org/10.1038/s41467-022-27980-y

VI. Figure 6) Le, Q.B., Nkonya, E., Mirzabaev, A. (2016). Biomass Productivity-Based Mapping of Global Land Degradation Hotspots. In: Nkonya, E., Mirzabaev, A., von Braun, J. (eds) Economics of Land Degradation and Improvement – A Global Assessment for Sustainable Development. Springer, Cham. https://doi.org/10.1007/978-3-319-19168-3_4

VII. figure 7) Azote for Stockholm Resilience Centre, based on analysis in Wang-Erlandsson et al 2022, https://www.stockholmresilience.org/research/planetary-boundaries.html

Author: Martin Roe