Climate Finance Innovation: The Value of Machine Learning

Andres Alonso-Robisco¹, Jose Manuel Carbo², Jose Manuel Marques³
Banco de España ¹, ², ³, Universidad Autónoma de Madrid ¹
andres.alonso@bde.es ¹, jose.carbo@bde.es ², manuel-m@bde.es ³

Abstract

Preventing the materialization of climate change is one of the main challenges of our time. The involvement of the financial sector is a fundamental pillar in this task, which has led to the emergence of a new field in the literature, climate finance. In turn, the use of Machine Learning (ML) as a tool to analyze climate finance is on the rise, due to the need to use big data to collect new climate-related information and model complex non-linear relationships. Considering the potential for the use of ML in climate finance and the proliferation of articles in this field, we propose a survey of the academic literature to assess how ML is enabling climate finance to scale up.

The contribution of this paper is threefold. First, we do a systematic search in three scientific databases to assemble a corpus of relevant studies. Using topic modelling (Latent Dirichlet Allocation) we uncover representative thematic clusters. This allows us to statistically identify three overarching research areas, and seven granular application domains where ML is playing a significant role in climate finance literature: physical risks (natural hazards, biodiversity and agricultural risk), transition risks (carbon markets and energy economics), and corporate and social responsibility (ESG factors & investing, and climate data). Secondly, we do an analysis highlighting publication trends, and a breakdown of ML methods applied; and thirdly, we provide a literature review of each topic, pointing out emerging research directions.

1. Introduction

The relationship between economic and environmental sciences has increasingly attracted the attention of economic researchers and professionals (Carvalho et al. 2016). The inception of the field, initially known as resource economics, is usually linked to the seminal work of Nobel Laureate William Nordhaus, who modelled the interactions between climate change and the economy. From there, more specifically on finance, early work on sustainability mainly addressed concerns on corporate governance and social investing (Capelle-Blancard & Monjon, 2012), not been until years later when more articles were produced on a broader set of problems related to climate. Interestingly, a change in the publication regime is observable from the year 2015 onwards (Malhotra & Thakur 2020), which might be associated with the signing of the Paris Agreement (UNFCCC, 2015), as suggested by Cunha et al. (2021). The agreement followed a warning by environmental scientists back in 2014 with the publication of the fifth assessment report (AR5), which defined the anthropogenic nature of climate change from unabated greenhouse gas emissions. This emergency call has been corroborated by the sixth assessment report AR6 (2022) which illustrates the lag still existing on policy implementations to effectively limit global temperature increase to well below 2°C by the end of the century. This scientific evidence is pushing the field of climate finance as a priority and triggering an impressive growth in publications. In fact, Liang & Renneboog (2021) note that despite its relatively late advent, the literature has been growing exponentially and evolving towards various topics.¹ As evidence, high quality economic journals now dedicate special issues to climate and sustainable finance topics.² This new prolific academic work has also been accompanied by international financial regulators and supervisors who have been actively working recently within and across institutions to scale up climate finance to develop a new financial architecture that properly incorporates and manages climate-related opportunities and risks. Just as an example, the European Central Bank set up in 2021 a dedicated ECB Climate Center and the Federal Reserve joined the Network for Greening the Financial System (NGFS) in late 2020.

¹ Notably, we reiterate that this sudden interest on the topic exploded only some years ago. As evidence, the three top finance journals (Journal of Finance, Journal of Financial Economics and Review of Financial Studies) did not publish a single article related to climate finance between January 1998 and June 2015, as indicated by Diaz-Rainey et al. (2017).

² For instance, the Review of Financial Studies, in March 2020. Also, thematic journals on environmental, climate and resource economics appear on top rankings like IDEAS/RePEc.
One characteristic of climate finance literature is how fragmented the research is. This is not only a bibliographic concern, as it also makes it difficult to join efforts from different academic profiles in order to develop specific research. In a literature review performed by Cunha et al. (2021) the authors conclude that it “makes it difficult to identify what constitutes the field and what differentiates it from traditional finance”, due to the poor theorization of the concept of “sustainability”, an opinion shared by several others like Capelle-Blancard & Monjon (2012), Zhang et al. (2019), Talan & Sharma (2019), Liang & Renneboog (2021) and Giglio et al. (2021). This calls for a precautionary need to define the scope of our survey on climate finance. We will rely on the definition provided by Giglio et al. (2021) as “the tools of financial economics designed for valuing and managing risk which can help society assess and respond to climate change”. Although we will use from now on the term climate finance, we acknowledge that three concepts are used indistinctively in the academic literature, namely green finance, climate finance and carbon finance (Zhang et al., 2019).  

Another characteristic of climate finance as a field is the difficulties researchers have to face in order to perform a solid empirical analysis. To name some of them: the growing access to climate data, though still of limited reliability, and the statistical complexity to model the non-linear behavior of climate change. These kind of problems create profound mathematical challenges for making inference about the real climate (Stephenson et al., 2012) and its relationship with the economy. In fact, Diaz-Rainey et al. (2017) conclude that methodological constraints could explain previous lack of climate finance research in top finance and business journals. Additionally, classical problems like the presence of endogeneity is cornerstone in climate finance, as the impact of climate on the economy is two-folded due to the existence of a feedback loop. This has been widely recognized by policy makers (European Commission, 2020), academicians (Gourdel et al. 2022) and financial supervisors (NGFS 2019, NGFS 2021). At the same time all these features justify the recourse to ML from researchers and experts as this technology is particularly well suited to deal with these problems. Taking into account the proliferation of articles in climate finance, the increased use of ML, and the fragmentation of the literature, in this article we propose a systematic review of studies that rely on this technology to solve climate finance problems. To face the challenge of heterogeneity of topics within the field, this survey leverages on Natural Language Processing (NLP), in particular we implement a Latent Dirichlet Allocation (LDA) model, to statistically uncover latent topics which we are then able to successfully identify as relevant application domains. To the best of our knowledge this is the first survey that thoroughly covers ML-based studies in climate finance, building a unique set of papers from different public databases, such as Web of Science, Google Scholar and Dimensions.ai. Notably, we make an effort to map which ML methods are mostly used by each climate finance topic, aiming to facilitate a profound understanding of how ML can enable climate finance to grow as a research field. This could be useful for future researchers interested on joining this academic debate, as well as policy makers looking for ways to better design climate finance instruments and policies. Indeed, the value of academic research in the overall innovation process has been widely investigated (see for instance Quatraro & Scandura 2019), and in climate finance this has been recently recognized in the last Conference of the Parts (COP26), where it has been stated that AI and ML can play a key role in important climate-related topics like prediction, mitigation, and adaptation, in ways we cannot afford to ignore (Clutton-Brock et al. 2021).

Importantly, our survey results support the relevance of ML as a driver of the publication trend in climate finance, a view that is starting to gain traction within economic and financial journals (e.g.: Musleh-Al-Sartawi et al. 2022). It is also aligned with the nascent concern from financial authorities on understanding the potential application of new technologies to resolve operational problems identified in climate finance. We refer for example to the G20-BIS Techsprint 2021, a race horse between private sector players leveraging technologies to solve a series of pre-identified problem statements (climate data collection, analysis of climate-related financial risks, and better connecting projects with investors). On this same effort, we can also highlight the global Fintech Hackcelerator for a greener financial system sponsored in 2021 by the Monetary Authority of Singapore, or the 2021 Green Fintech Challenge, hosted by the Federal Conduct Authority in UK. Significantly, with a longer term view, the Bank of International Settlements has created a series of Innovation Hubs (BISIH) worldwide, and a Network (BISIN) who are experimenting and monitoring new developments in climate finance innovation. Finally, the success of ML applied to climate finance issues is also corroborated by a new wave of projects and market-driven solutions which are flourishing in the private sector, giving birth to a new market segment currently labelled as “green fintechs” (Machiavello & Siri 2020).

The paper is organized as follows. Section 2 explores the role of ML in climate finance. Section 3 explains the methodology of the survey based on topic modelling. Section 4

---

3 Similarly, we will leave out of our scope any work not touching upon climate change, and exclusively focusing on social topics, like corporate governance, impact investing, social investing and financial inclusion, which would fall under the label of sustainable finance. Though, a limita-

4 Results shown in the Appendix.
details the findings from the clustering, and analysis of publications. Section 5 includes some analysis on publication trends, and Section 6 concludes.

2. The role of Machine Learning in Climate Finance

According to Athey (2018), ML is a field that develops algorithms designed to be applied to datasets, with the main areas of focus being prediction, classification, and clustering or data processing (e.g.: dimensionality reduction). While conventional statistical and econometric techniques, such as regression, often work well, there are idiosyncratic methodological problems that may benefit from using different tools. This is particularly relevant in climate-related issues.

First, the usual large size of the datasets involved in climate finance may require more powerful statistical manipulation tools. In recent years, the quantity and granularity of economic data in general has improved dramatically. The good news is that the sudden explosion of micro-level datasets offers an unparalleled insight into the inner workings of the economy and financial systems. The bad news is that datasets are increasingly more complex to deal with (López de Prado, 2019). As an example with climate finance implications we can point just to the high variance among the temperature predictions of the 20 global climate models, from various laboratories around the world, that inform the Intergovernmental Panel on Climate Change (IPCC), with data for over 100 years (Monteleoni et al. 2011). In fact, some of the most interesting datasets in climate finance are not only high-dimensional, but also unstructured, like news articles, voice recordings or satellite images, which along with the complexity of the phenomena they measure, means that many of these datasets are beyond the grasp of econometric analysis.

Second, big datasets may allow for more flexible relationships between the variables than simple linear models. It has been largely recognized that ML techniques such as decision trees, support vector machines, neural nets, deep learning, and so on, may allow for more effective ways to model complex financial and economic relationships (Varian 2014, Athey 2018, Athey & Imbens 2019). The key advantage and one common feature of many ML methods is that they use data driven model selection, treating the data-generating process (DGP) as unknown, allowing researches to deal with large datasets without imposing restrictive assumptions. On the other hand, as described by Breiman (2001), traditional model-driven statistical community assumes that the data are generated by a given stochastic process, being able to better understand the relationship between the variables.

As very illustratively explained by Huntingford et al. (2019), and Castle & Hendry (2022), shared characteristics of financial and climate time series make ML tools appropriate for studying many aspects of observational climate-change data and its economic impact. For instance, greenhouse gas emissions are a major cause of climate change as they accumulate in the atmosphere. As these emissions are currently mainly due to economic activity, financial and climate time series have commonalities, including considerable inertia, stochastic trends, possible nonlinearities, omitted variables and abrupt distributional shifts. Moreover, both disciplines lack complete knowledge of their respective DGPs, so data-driven model search allowing for shifting distributions is important; and ML offers a rigorous route to analyzing such complex data. The appeal of ML is that it manages to uncover generalizable patterns. In fact, the success of ML is largely due to its ability to discover complex relationships that were not specified in advance. It manages to fit complex and very flexible functional forms to the data without simply overfitting, that work well out-of-sample (Mullainathan & Spiess, 2017).

Importantly in climate finance, ML offers us the opportunity to explain relationships that have the potential for huge societal impact (Hoepner et al. 2021). Indeed, the effects of climate change are increasingly visible. Storms, droughts, fires, and flooding have become stronger and more frequent. Global ecosystems are changing, including the natural resources and agriculture on which humanity depends. Yet, year after year, these emissions rise, giving only a pause during Covid-19 lockdown. In the well-known “Tragedy of the Horizons”, Mark Carney (2015) showed us that the environmental impact of climate change translates into substantial financial risks to global assets measured in the trillions of dollars. However, it is hard to forecast where, how, or when climate change will impact the stock price of a given company, or even the debt of an entire country. Financial short-termism fails to incentivize the prediction of medium or long-term risks, which include most climate-change-related exposures such as physical impact on assets. As we will see, ML can help us to close this “intertemporal” gap. A very illustrative example is given by researchers from the Quebec AI Institute (2021), who warned during the last COP26 that preventing climate-related catastrophic consequences will require changes in both policy-making and individual behaviors. However, many cognitive biases (like abstraction and myopic term discount) might prevent us from taking action today. To tackle this market failure, they developed “This Climate does not Exist”, a research project that harness ML (in particular Generative Adversarial Networks, or GANs) to create images of personalized climate impacts which will be especially powerful in overcoming the barriers to action and raising climate change awareness.
On measuring climate awareness, ML and alternative data are playing a key role themselves. For instance, Sluban et al. (2015), Cody et al. (2015) and Shangguan et al. (2021), employ different ML algorithms to analyze the sentiment toward selected topics related to the environmental issues in Twitter. Drawing on large-scale computational data and ML methods, Farrell (2016) shows that media organizations with corporate funding were more likely to have written and disseminated texts meant to polarize the climate change debate. On the same front, Stecula & Merkley (2019) use a Support Vector Machine (SVM) to identify economic, conservative ideological, and uncertainty frames in the press coverage of climate change in some of the most influential US media sources. Lynam (2016) explores social adaptation to climate change from a set of microrattles collected in Australia, using topic modeling and Bayesian networks, and Gaur et al. (2021) evaluates the inclination of Asian youth regarding the achievement of the SDGs, building on a Random Forest model to capture their opinions about a sustainable future. But what about financial markets’ view climate change? Interestingly, to answer this question Schlenker & Taylor (2019) rely on ML as well, to pick the optimal model to estimate the market expectation on climate change. They analyze prices of financial products whose payouts are tied to future weather outcomes, and their results suggest that trends in weather markets follow climate model predictions, and are not based on shorter-term variation in observed weather station data. This promising result would indicate that when money is at stake, agents are accurately anticipating warming trends in line with the scientific consensus of climate models.

But the set of topics in climate finance where ML is being utilized is much broader. Rolnick et al. (2022) show how ML can contribute, for instance, in climate investment, applying deep learning both for tilting portfolio selection towards low carbon emitting corporates, and investment timing. In fact, as concluded by the authors, this climate-aligned investment strategy is creating major shifts in certain sectors of the market towards renewable energy alternatives, which are seen as having a greater growth potential than traditional fossil fuels. This is another example of the high impact of climate-related problems. Due to dependencies from several nations on Russian oil and gas, the green transition has gained a further sense of emergency, having its implications on the future regulation of energy markets (e.g.: Plan RePowerEurope). Furthermore, we could further elaborate on the overlapping issue between green public policies and digitalization. For instance, Gailhofer et al. (2021) specifically discuss about the role of AI in the European Green Deal, Bag et al. (2021) study the role of public institutions on the adoption of big data analytics and AI technology, and how this affects sustainable manufacturing and circular economy, and Plakandaras et al. (2019) use ML techniques to model climate change as a geopolitical risk, forecasting its impact on several financial assets.

As a conclusion, the emerging use of AI and ML is disrupting and transforming the financial industry (Wall, 2018). Climate finance is a particular area where innovation is growing fast and having big impact, as acknowledged by academics, policy makers, and market participants. As an example, in a position paper Kaack et al. (2020) hope that recent breakthroughs in ML can help us get closer to achieving the UN SDGs, and Kumar et al. (2021) think that new-age technologies applied to sustainability can make significant contributions to the green transition. Both Al-Sartawi et al. (2021) and Avgouleas (2021) suggest that cutting-edge financial technology encompassing AI, ML and blockchain can be critical in terms of boosting sustainable finance. And for Inampudi & Macpherson (2020) there is a great potential for AI to contribute towards global economic activity, especially ESG investing. In fact, the digitalization of climate finance has led to the birth of a FinTech sector that comprises technology-backed innovative business models for finance, something that has been given the name of “Green Fintech” (see GDFA 2022 for a taxonomy devoted to classify market-driven green fintech business solutions).

Finally, we feel responsibly obliged to bring to this discussion the other side of the impact of ML on climate change, as well. New technologies do not only bring us opportunities. Kaack et al. (2020) explain ways in which AI and ML can be detrimental to efforts addressing climate change, warning of those uses that might harm our planet. AI or AI-driven technologies can become pollutants and net emitters of greenhouse emissions, depending on the types of applications and the circumstances of their deployment. For example, remote sensing algorithms for satellite image analysis can be used to gather information on agricultural productivity, but can also be used to accelerate oil and gas exploration. Self-driving cars can make driving more efficient, but they could also increase the amount people drive. And finally, ML include computationally expensive programming, which is an energy intensive activity. This final concern has minted the term “Green AI”, referring to responsible and low carbon intensive coding and good practices relating the training and deployment of complex algorithms in the academic industry (e.g.: Strube et al. 2019, or Hirschovich et al. 2022). We include a dedicated literature review on this topic in the Appendix.

3. Methodology

We adopt and implement the Scientific Procedures and Rationales for Systematic Literature Reviews (SPAR-4-SLR)
Our final collection of documents adds up to 217 research articles, from which we extract the abstracts, which will comprise the sample of texts (corpus) in our study. Our goal will be to discover the hidden or latent (unobservable) topics in the corpus of documents (observable), using a ML technique, Latent Dirichlet Allocation (Blei et al. 2003). This will help us understand documents analyzing the presence of words. Often the term “topic” is used in a technical, statistical sense, but ultimately the last phase of any topic modeling approach involves expert inspection to uncover through a more usual theme that aligns with each topic, allowing to name each of them with a more meaningful name. In addition, we aim to rank the topics according to their prevalence (Sievert & Shirley, 2014), which we find to be a convenient visualization tool for the exploration and presentations of the topics.

Data collection
To assemble the corpus of articles on ML-based climate finance, we identified relevant keywords relating to climate finance from a preliminary assessment of literature reviews on both sustainable (carbon, or green) finance, energy economics and ML in finance (i.e.: Kumar et al. 2021, Ghodsi et al. 2019, and Aziz et al. 2019). Following the identification of these words in climate finance and ML (this led to a combination of 15 keywords) we conducted a search for articles using an advanced search string in the category ALL (“article title, abstract, and keywords”), and AB (“abstract only”) on Google Scholar, Web of Science, and Dimensions.ai, as shown in Expression 1. The start date was selected to be 1st January 1999, until the present day, being the last update as of April 22nd, 2022.

Expression 1

\[
\text{ALL}=\text{“climate change” OR “ESG” OR “sustainable finance” OR “green finance” OR “climate finance”} \text{ AND AB = (finance OR “financial market”* OR bond* OR investment* OR corporate* OR funding OR financing) AND ALL= (“lasso” OR “random forest”* OR “extreme gradient” OR “xgboost” OR CART OR “deep learning” OR “neural network” OR “machine learning”)}
\]

The data was collected The data was collected using a “Human-In-The-Loop” (HIL) approach. It consists of proceeding to a purely automated data collection with an ex-post validation based on the field expertise. For instance, a total of 45 search pages (showing 10 items each) were screened in Google Scholar and the process of checking potential duplicates between different databases was performed by an expert using the software OneNote. Contrary to other reviews, we aim to focus on a narrow definition of ML in climate finance. This means our results should be familiar to economists and not relying too heavily on environmental or engineering science with no connection of the research question or conclusion to an economic of finance theme or discourse. It is important to highlight that our approach, incorporating a screening phase in Google Scholar, which allows for a richer understanding of a research field that is growing so fast, and therefore so much relevant research is still in working paper status, waiting to be published by peer-reviewed journals, and therefore do not appear in results from more standardized databases like WoS or DAI yet.

Topic modelling
Topic modeling assumes a person approaches writing a document with a collection of topics in mind and the words chosen will represent this topic mixture. For instance, a climate finance researcher applying ML to solve a problem will, for example, write a paper with a topic mixture of 50% climate change, 30% finance, and 20% ML modelling. The key task for the topic modeling researcher is therefore to reverse engineer the latent topics from the observed words. Currently, the most widely accepted approach for topic modeling is Latent Dirichlet Allocation (LDA) developed by Blei et al. (2003). The key practical advantage of LDA is that it allows documents to be a mixture of different topics, while topics are presented as a mixture of words. This fits the reality observed in climate finance studies, since different topics can partially overlap within a document. We apply the Gensim implementation of LDA in Python (Rehurek and Sojka 2010). The procedure for extracting the topics consist of a variety of steps required for training, tuning, and applying the resulting LDA model to the corpus. We briefly describe the most important ones:

A necessary first step in topic modeling is processing the corpus of documents by tokenizing each document into a collection of their individual words where order is unimportant (i.e.: each document is treated as a “bag of words”). Then, “stop words” that have no topic context (such as “and”, “of”, “the”), are removed, as well as common terms that are highly repeated in the corpus, which we identify because they appear in more than half of the documents, or rare

\[\text{As a robustness check we verified that all the studies tagged as “climate finance and economics” in the expert network hosted in https://www.climatechange.ai/ were included.}\]

\[\text{This was actually a drawback we saw from other related literature reviews like Warin & Stojkov (2021), or Kumar et al. (2021).}\]
terms for which we set a threshold of being in less than two documents. We deem that both categories of terms contain little meaning to contribute to a relevant topic. Remaining words in a document are stemmed to generate the words’ root, and accurately capture unique terms usage. This means suffixes are removed to create common stem terms, e.g.: finance, financial and finances might be reduced to the common “finance” root. In theory, a token can have any number of words (usually monograms are used, but we could have bi- and trigrams). For simplicity, we keep our analysis to single word tokens as we find that it allows us to easily label the topics at the final stage.

After processing the data, we count with $D$ documents that together contain $N$ unique tokens that we can represent by an $N \times D$ matrix $W$ with entries $w_{n,d}$ that are the number of occurrences of token $n$ in document $d$. Thus, the total number of tokens in document $d$ is $N_d = \sum_{n=1}^{N} w_{n,d}$. The LDA model consists of two matrices, $\beta(N \times K)$ and $\theta(K \times D)$, where $K$ is the total number of topics. For topic $k$, the vector $\beta_k$ contains the $N$ token weights, which act as the probabilities $P(n|k)$ that the token $n$ contribute to a document’s bag of words, conditional on the topic $k$ contributing to the document. That is, $P(n|k) = \beta_{n,k}$, i.e.: the weight of token $n$ in topic $k$. Therefore, $\sum_k^{N} \beta_{n,k} = 1$. For document $d$, the vector $\theta_d$ contains the $K$ topic weights – which act as the probabilities $P(k|d)$ that the topic $k$ appear in the document. Thus, $P(k|d) = \theta_{kd}$, i.e.: the weight of topic $k$ in document $d$. Similarly, $\sum_k^{K} \theta_{k,d} = 1$. When these probabilities are significant, we may say that a topic $k$ is relevant in document $d$. Finally, this setting allows us to decompose in Equation 1 the probability of a token $n$ in a document $d$ as (Hofmann 2001):

$$
\text{Eq. 1} \quad P(n|d) = \sum_{k=1}^{K} P(n|k) \cdot P(k|d) = \sum_{k=1}^{K} \beta_{n,k} \cdot \theta_{k,d}
$$

Topic modeling involves reducing the dimensions of these matrices to end up with the same number of rows (documents) but a restricted number of columns which represent the topics. To this purpose LDA assumes a particular Dirichlet distribution that can be used to produce probability vectors $\beta_k$ and $\theta_d$, that allow an assumption to be made about how topics are distributed across tokens and documents. Using two external inputs, $\alpha$ and $\beta$ as Dirichlet priors, we can determine the generative process in the LDA. $\alpha$ is a parameter that determines $\theta_d$ or per-document topic distribution, and $\beta$ is a parameter that determines $\beta_k$ or per-topic token distribution. The LDA posteriors are a result of the trade-off between two inherently conflicting goals. Firstly, that only a relatively small number of topics are expected in a well-written document, and secondly that only high probability should be assigned to a small number of tokens that belong to highly informative topics. The trade-off exists because if we assign, for instance, a single token to a single document, thus succeeding at the first goal, the second goal becomes difficult to achieve because all tokens in the document must have a relatively high probability of belonging to that topic. The estimation of the LDA model requires a Bayesian updating from its initial semi-random allocation of topics to tokens and documents, to converge to a probabilistic distribution of topics across documents. Technically, the process will be completed when we find matrices $\beta_{n,k}$ and $\theta_{k,d}$ that most likely have produced the observed data $W$.

4. Survey Results

As we mentioned, LDA becomes a useful approach to cluster similar documents together from a large disparate literature, as it is the case of ML-based climate finance. To select the number of topics for our final model, multiple models with different topic numbers were produced and relevance scores were compared, following Equation 2.

A challenge with topic modeling is that topics that make ML-sense do not necessarily make human sense. Therefore, in order to label the resulting topics, we do a qualitative check with human expert judgement to ensure that the words determined for each topic make sense within the existing climate finance literature. When the LDA model is estimated, we inspect the topics in three ways: we look at the tokens with the highest probability per topic ($\beta_k$); we sample $d = 20$ documents and check whether the highest probability $\theta_{k,d}$ of each document $d$ belonging to a topic $k$ matches the thematic area identified by a human expert in advance (who read the abstract)$^{10}$; and finally we look at the tokens ranked according to topic relevance as defined by Sievert & Shirley (2014). The relevance $r$ of token $n$ to topic $k$, given a tuning parameter $\lambda$ is given in by:

$$
\text{Eq. 2} \quad r(n,k|\lambda) = \lambda \log \beta_{n,k} + (1 - \lambda) \log \frac{\beta_{n,k}}{\sum_{k=1}^{K} \beta_{n,k}}
$$

Where the term $\log \frac{\beta_{n,k}}{\sum_{k=1}^{K} \beta_{n,k}}$ is called token’s lift. The higher the marginal probability of token $n$ over the corpus, the higher is its lift and the more exclusive a token is for a topic. With $\lambda = 1$, tokens of top relevance equals the top words, $^{10}$ All results present herein pass this test, with a threshold of at least 50% success rate.
even if these do not show up exclusively in that particular topic. With \( \lambda = 0 \), tokens of top relevance are the ones exclusive to the given topic. By varying \( \lambda \in (0,1) \) and studying the different resulting ranking of tokens, we get a good understanding of the words that contribute to a topic. Following the recommendation of Sievert & Shirley (2014) we fix \( \lambda = 0.66 \) in order to label them with an economic meaningful name.\(^{11}\)

As climate finance is a very fragmented topic, we are interested on understanding first which might be the big overarching areas of research that clearly appear in the literature, before going any further in granularity. Therefore, we start by estimating a model with three topics, because we want to grasp a big picture of overarching areas that drive the research agenda in ML-based climate finance. In Table 1 we show the estimated vector \( \beta_k \) per topic, as extracted from the LDA model. After inspection, we label each topic, being able to identify three overarching thematic areas. Topic 1 could be matched with transition risk, including words related to carbon emissions, energy economics, and buildings efficiency. Then, we name topic 2 as physical risk, including references to flood (natural disasters), agricultural risk, climate change, crop performance and forests. Finally, inspecting topic 3 we observe words linked to corporate & social responsibility, like ESG factors, (responsible) investing, companies’ performance, sustainability and risk. We show these results in the Appendix.

Once we understand that ML is used in three big reasonable areas of research of climate finance we wonder whether we can locate narrower topics, in order to provide better field knowledge for future researchers. There is no easy way to find the optimal number of topics. The most commonly used statistical measures to determine an appropriate number of topics are the coherence score (Röder, 2014) and the statistical perplexity during topic modelling (Blei, 2003). While increasing the number of topics usually improves these statistical measures during topic modelling, we must at the same time account for a higher computational cost of training the model as the number of topics increase, and more importantly, the complexity for a human to discern the economic meaning of more topics will also increase. In the Appendix Figure 12, we plot the coherence score and the latency of training during the estimation of up to 40 topics, and we evidence that the coherence finds a stable trend close to 15 topics. Following Zhao et al. (2015) we also compute the rate of change in perplexity (RCP), as a more suitable selector of the number of topics (see Appendix, Figure 13). This plot converges to a stable level around 5 to 10 topics. Observing this, we decide to estimate our LDA model with 10 topics, knowing that the ability of a human to label the economic meaning of the topics will be more challenging. At the end, we are able to label a total of 7 comprehensive and economically reasonable topics, having to discard 3 of them. We decide to stop here, as more granularity offered no further insights to human experts inspecting the tokens in too small topics. In Table 2 we show the newly estimated vector \( \beta_k \) per topic (see Appendix).

By inspecting these keywords, we can again initially label each topic, resulting this process in the following more granular research areas in climate finance that rely on ML-methods: (i) natural hazards, (ii) biodiversity, (iii) carbon markets, (iv) agricultural risk, (v) ESG factors & investing, (vi) energy economics, and (vii) climate data. We discard three topics because we find that their composition is either mainly comprised of methodological terms (e.g.: in topics 1 and 3 we encounter tokens like “activ”, “correl”, “signific”, “algorithm”, “term”, “price”, “differ”, etc.) or repetitive with other topics (e.g.: in topic 5 we find concepts related to carbon markets like “emiss”, “carbon” and “soil”, but mingled with low relevant tokens like “stud”, “result” and “forecast”. This, together with an investigation of the relevance of this topic using \( \lambda=0.66 \) makes us discard it). At this stage we are interested again in checking the inter-topic distance, to confirm the link with our previous classification and how the three overarching thematic areas might breakdown in these reasonable more granular research themes. In Figure 4 we plot the visualization of the new clustering in 7 meaningful topics.

Figure 1. Visualization of topic 9 (Energy economics)

From this previous picture, we first observe that topics 6 (carbon markets) and 9 (energy economics) are indeed very close of each other, and might reflect a breakdown of the overarching area labeled before as Transition risk, as it is efficiency (Hoffman et al. 2010). In order not to stop at a local optimum we use a high enough number of iterations. Finding 3 topics is an easier task for the LDA model, so we find a stable result using only 100 passes. However, in order to find a higher number of (meaningful) topics we will need 40,000.

\(^{11}\) Importantly, running a LDA model implies an iterative process, to find the optimal configuration. The more passes, increases the fitness of the model. In our case, the Gensim implementation, based on a Bayesian approach, finds the best configuration of the model automatically as well as several settings related to numerical.
suggested by the similarity of their principal components. Then, topics 8 (ESG factors & investing) and 10 (Climate data), again fall nearby one from each other, and reflect therefore correctly a breakdown of the previously named area Corporate & Social Responsibility. Finally, we discover that topic 7 (agricultural risk), topic 2 (Natural hazards) and topic 4 (Biodiversity) cover a similar theme, being all of them connected to the previous area named Physical risk.

Therefore, we successfully arrive after inspection of the relevance scores of key tokens per topic to a meaningful understanding of the concepts covered by each topic. For instance, using as example Figure 4 for Topic 9, in the right hand side panel, we find highly ranked (nearly) exclusive terms like “energi”, “emiss”, “carbon”, “ghg” or “greenhous”, as well as overlapping terms like “predict”, “carbon”, and “build”. Varying the values of λ, we can easily label this topic as Energy economics, understanding this as a cluster of research paper dealing with ML to solve problems related to GHG emissions, air pollution, carbon price, energy forecasting, energy consumption and buildings efficiency. For further reference we leave in the Appendix the visualization of the remaining topics, being able to confirm that the labeling makes economic sense after inspection of the respective relevance rankings, which also we must highlight that allow us to fine-tune the final name of each topic in further detail. In the following Section we leverage on the granular classification of research areas to study key publication metrics in order to understanding potential knowledge gaps, emerging areas of interest and key ML methods that are being utilized.

5. Publication Trends and Analysis

Using the classification into topics that we found in the LDA analysis (both the classification into three topics and the more granular classification into seven topics) we show in the Appendix (Table 1) a descriptive summary of key statistics of the corpus under scrutiny. From a total of 217 unique documents, Physical risks, Transition risks and CSR capture a similar share of total publications. However, Physical risk seems to be a more mature research area as the majority of publications are in peer-reviewed journals. This contrasts with other areas that seem to be emerging and relying still more on working paper format, especially two, Climate data, where more than half of the research articles gathered are still in not published in a journal, and ESG factors & investing, where notably close to half of the documents belong to this class. Further analysis is provided in the Appendix, including a breakdown of ML methods used per topic.

From our results, we are able to extract some stylized conclusions. First we observe that currently ML is applied for a majority of topics related to climate change in finance. For instance, we identify relevant studies covering five out of the seven topics listed in Kumar et al. (2021), which could serve as a benchmark survey describing the field of sustainable finance. Then, we show evidence on how climate finance researchers are broadening the use of ML. From being initially applied to solve physical risks problems, like weather and natural hazards forecasting, and issues related to energy economics, currently a relevant number of studies are using it for responsible investing, ESG factors and measuring corporate’s compliance with climate data regulatory disclosures, although these latter areas are less mature, as shown by the ratios of peer-reviewed publications versus working papers. Interestingly, some ML models outstand within each field of interest. Overall, Random forests and Artificial Neural Networks are the mostly used ones, but for instance, climate data is heavily relying on Natural Language Processing, and Corporate & Social Responsibility is still heavily relying on more traditional tools like Lasso/ Ridge and Elastic net regularization in multiple types of regressions.

6. Conclusion

We aim to shed some light on the value of ML within climate finance, in order to understand its potential to drive innovative work in this knowledge area. To this purpose we assemble a corpus of relevant articles and we estimate a Latent Dirichlet Allocation model to uncover latent topics in the literature, finding three overarching areas and seven granular application domains which we are able to label with economic meaning that significantly describe where ML is being used within climate finance. To the best of our knowledge this is the first study that relies on Natural Language Processing to survey this highly heterogeneous research field, offering academics, market experts and policy makers a means to assess emerging topics, and promising thematic areas. We hope this will enable a profound knowledge of the field, aiding climate finance to scale up in order to become mainstream finance in the near future.

As a bottom line, climate finance literature has been growing fast, and we have been able to demonstrate the importance of ML in this field. We uncover up to seven research topics that are coherent with current sustainable finance literature reviews, and illustrate the areas where ML is adding more value. For instance, climate data seems to be a novel area that is arising thanks to ML. We also identify topics like physical risk that remains mainly covered by environmental journals, while economic journals seem prioritize research on ESG and carbon markets, having therefore to acknowledge that the relevance of climate finance is still a work in progress in the top academic arena in economics.
References


through the application of machine learning to satellite images. In 2020 WORLD BANK CONFERENCE ON LAND AND POVERTY.


License: CC BY 3.0 IGO.


Citterio, A. The role of ESG in predicting bank financial distress: cross-country evidence.


Green Digital Finance Alliance (2022). A Green Fintech Taxonomy and Data Landscaping


models for medium-term forecasting of water inflow when planning electricity generation under climate change. Energy Reports, 8, 439-447.


International Conference on Smart City Applications (pp. 1-7).


Yan, J., & Meng, B. (2021). Stock Market Reactions to Firms’ First Publication of Sustainability Reports: Evidence from multiple industries in Nordic countries, the United States, and China.


in China: An empirical study of energy service companies. Journal of Cleaner Production, 266, 121395.


Appendix
Publication Trends and Analysis

Table 1. Descriptive statistics of the corpus.

<table>
<thead>
<tr>
<th>Research Area</th>
<th>Count</th>
<th>Mean</th>
<th>Median</th>
<th>Max</th>
<th>Min</th>
<th>SD</th>
<th>Median Abs Dev</th>
<th>Skew</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical Data</td>
<td>234</td>
<td>7.2</td>
<td>5</td>
<td>10</td>
<td>0</td>
<td>3</td>
<td>1.96</td>
<td>0.61</td>
<td>0.45</td>
</tr>
<tr>
<td>Transition Risks</td>
<td>284</td>
<td>7.8</td>
<td>7</td>
<td>12</td>
<td>0</td>
<td>3</td>
<td>1.03</td>
<td>0.23</td>
<td>0.12</td>
</tr>
<tr>
<td>Corporate &amp; Social Responsibility</td>
<td>320</td>
<td>9.0</td>
<td>7</td>
<td>15</td>
<td>0</td>
<td>5</td>
<td>1.65</td>
<td>0.56</td>
<td>0.41</td>
</tr>
<tr>
<td>Total</td>
<td>838</td>
<td>7.5</td>
<td>5</td>
<td>15</td>
<td>0</td>
<td>4</td>
<td>1.63</td>
<td>0.53</td>
<td>0.39</td>
</tr>
</tbody>
</table>

The emerging importance of some topics can be also concluded from Figure 4 which shows the cumulative number of publications per year and topic. We highlight that Energy economics (gray line) is a highly covered area, showing sharp growth rates as well Carbon markets (yellow line) and ESG factors (orange line). Agricultural and extreme weather events is possibly the area with the most stable growth rate, indicating the reliance of researchers and experts on this application domain of ML methods during the last years.

Out of those publications that were published in journal format (totaling 136), we are able to classify them per type of science. We identify publications of ML-based climate finance in very heterogeneous knowledge domains, like journals from environmental sciences, computer sciences, or economics and finance journals. In Figure 5 we plot this breakdown, concluding that Economic and Finance journals still pay more attention to topics related to CSR and Transition risks, lagging behind other scientific journals that publish more work on Physical risk and its socio-economic impact.

Finally, we are interested in investigating which are the most used ML models per topic, in order to provide valuable insights to experts willing to look into new fields where this tools are been successfully useful. In Figure 6 and 7 we show the breakdown for Physical and Transition risks. Very interestingly in Physical risk we appreciate a strong usage of image recognition tools, usually associated with the need to handle newly available (unstructured) data from remote sensing, text, and satellites. Following this method both Random forests and Artificial Neural Networks are widely used in this field of research. However, in Transition risks, Artificial Neural Networks dominate within our subset of documents, but can highlight that a relevant share of studies in this domain still use more traditional techniques like Lasso and other penalized-type regressions.
In Figure 8 we show the models predominantly used in CSR. Two facts are important to highlight, first that the ample dispersion of methods used in CSR; and secondly, the strong reliance in Natural Language Processing tools, clearly strongly used by studies focusing on how to extract and report climatic data.

To conclude, we extract one further interesting insight. In all the three overarching areas of research post-hoc interpretability techniques are starting to be used, gaining a relevant share in all these application domains, in line with recent advances in the academic field of Explainable AI. This trend goes in parallel with its gaining traction in several other areas of mainstream finance like risk management, both at academic level (see Albanesi & Vamossy 2019, or Watcher et al. 2017, for example) and within financial supervisors and regulators, as illustrated in IIF (2020), EBA (2021), Akinwumi et al. (2022), Dupont et al (2022), Bafin (2022), or Alonso & Carbó (2022).

Methodology

In Figure 9 we show the models predominantly used in Physical risks and in Transition risks. Figure 10 shows the rate of change in perplexity and latency of training LDA.
**Green AI**

Recently artificial intelligence has encountered such dramatic progress that it is seen as a tool of choice to solve environmental issues, such as greenhouse gas emissions (GHG). At the same time the ML researchers began to realize that training models with more and more parameters required a lot of energy and as a consequence GHG emissions, questioning the complete environmental impacts of AI methods for the environment (see Schwartz et al. 2020). Based on this concern, Ligozat et al. (2021) propose to study the possible negative impact of AI systems often presented as a solution to climate change, presenting different methodologies used to assess this impacts, in particular life cycle assessment. For instance, recent advances in large Transformer models have been taking seriously into consideration their environmental footprint at the time of designing and developing the models (see Zhang et al. 2022b).

However, as we are seeing in our study, a large variety of ML methods are used in Climate Finance, making sense to extend the concern on the environmental footprint of ML more broadly. Strubell et al. (2019) in a pioneer paper estimated the consumption of large NLP models, comparing it in CO2 equivalents with illustrative general life examples. They conclude that training a big Transformer with neural architecture search can emit up to six times what a car produces (including fuel) in its lifetime. Therefore, the authors recommend to grant researches equitable access to computation resources, and suggest to prioritize computationally efficient hardware and algorithms. In Strubell et al. (2021) they extend their work to modern language models like BERT, or GPT-2.

Overall, a common conclusion is that we need accurate reporting of energy and carbon usage. It is essential for understanding the potential climate impacts of ML research to incentivize responsible research. To this purpose, Henderson et al. (2021) introduce a framework that makes this easier by providing a simple interface for tracking ML models’ real-time energy consumption and carbon emissions, making carbon accounting easier. Lacoste et al. (2019) present as well a Machine Learning Emissions Calculator as a tool for researches to better understand the environmental impact of training their models. In a position paper Schwartz et al. (2020) advocates a practical solution by making efficiency an evaluation criterion for research alongside accuracy and related measures, like Hershcovich et al. (2022) who propose a climate performance model card with the primary purpose of being practically usable with only limited information about experiments and the underlying computer hardware, in order to increase awareness about the environmental impact of NLP research.

A big challenge remains on new methods being currently develop to achieve a trustworthy and scalable ML. For in-
Clustering results: 3 thematic areas

Figure 11. Visualization of Topic 2 (physical risk)

Figure 12. Visualization of Topic 2 (physical risk)

Figure 13. Visualization of Topic 3 (corporate & social responsibility)

Clustering results: 7 granular application domains

Figure 14. Visualization of topic 6 (carbon markets)

Figure 15. Visualization of topic 8 (ESG factors & investing)

Figure 16. Visualization of topic 10 (Climate data)
Figure 17. Visualization of topic 4 (Biodiversity)

Figure 18. Visualization of topic 2 (Natural hazards)

Figure 19. Visualization of topic 7 (Agricultural risk)