Regulatory benchmarking by machine learning: The case of climate resilience in electric utilities

by

Beichen Lyu

B.S. Environmental and Natural Resources Engineering & Computer Science, Purdue University (2019)

M.S. Computer Science, University of California, San Diego (2021)

Submitted to the Institute for Data, Systems, and Society
in partial fulfillment of the requirements for the degree of

Master of Science in Technology and Policy
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
September 2023

Authored by: Beichen Lyu
Institute for Data, Systems, and Society
August 25, 2023

Certified by: Christopher R. Knittel
George P. Shultz Professor of Applied Economics
MIT Sloan School of Management
Thesis Supervisor

Accepted by: Frank R. Field III
Senior Research Engineer, Sociotechnical Systems Research Center
Interim Director, Technology and Policy Program

Regulatory benchmarking by machine learning: The case of climate resilience in electric utilities

by

Beichen Lyu

Submitted to the Institute for Data, Systems, and Society
on August 25, 2023, in partial fulfillment of the
requirements for the degree of
Master of Science in Technology and Policy

Abstract

Regulatory targets are becoming increasingly complex to benchmark, including electric utilities' climate resilience ("utility resilience") that are non-linear and high-dimensional. Meanwhile, machine learning (ML) models have been developed, and continue to be developed, with desirable properties such as local optimality and data compression. To explore the synergy between ML and benchmarking, we review and discuss the literature from both sides in the context of government regulation.

Then we dive into a case study of utility resilience, where the dual complexities of the climate and power systems converge, with climate impacts that are likely to harm resilience and increase risks. However, these complicated and changing climate impacts are overlooked in the current regulations of utility resilience [30]. We examine how benchmarking could be applied to fill this regulatory gap through performance incentive mechanisms and elaborate on the political-economic implications, both advantages and potential pitfalls, of its application.

With these theoretical understandings, we experiment with benchmarking weather-related power outages in New England, US between 2010-2021. We propose a data regime by combining station-level weather data with district-level outage data, as well as a baseline model using ridge regression. We also deploy our model through an online portal, as well as discuss its limitations on long-tail distributed outage and weather data. Our studies could inform future ML-based benchmarking for regulatory uses, particularly over utility resilience, that balances accuracy, accessibility, and applicability.

Thesis Supervisor: Christopher R. Knittel
Title: George P. Shultz Professor of Applied Economics
MIT Sloan School of Management

Acknowledgments

to

- Chris, for being my "benchmark" of a thinker, a communicator, and a faculty advisor;

- Dan, Rob from National Grid, who provided funding, data, expertise, and trust;
- Steve, Daniel from Earth Networks, who offered lightning data, and encouragement;
- All CEEPR colleagues, especially Tony and Sanjam on assisting this project;

- TPP, for enabling this life-changing journey, and for the best Administrator, Barb;
- All TPP classmates, for their friendliness, including Alexa's LaTeX template;

- Frank, for his IDS.411 that rewires my tech brain, and for his genuine advice;
- Noelle, for her IDS.410 that inspired my policy interest, and for her open mind;
- Ken, for his IDS.412 that unifies my poli-econ knowledge, and for the kind TA, Diana;
- David, for his 11.381 that stretches my tech-policy thinking, and for his warm heart;
- All instructors, esp. Profs. Angrist, Solomon, for growing my intellectual aesthetic;

- Unsung open-source developers of Pandas, Sklearn, PyTorch, statsmodels...

- And my family, their unconditioned love, unmatched humanity, unbounded vision.

Section 2.1, subsection 2.2.2, chapter 3, and section 3.1 re-used part of my in-class essays from 11.381, IDS.410, IDS.412, and IDS.411, respectively. None of these essays were published, and I am their sole author with valuable feedback from class instructors or TA.

The contents are those of the author(s) and do not necessarily represent the official views of, nor an endorsement, by National Grid or Earth Networks.

All errors mine.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

How do we make electric utilities more resilient under climate change? Maintaining a resilient electric utility is critical not only because electricity empowers economic activities, but also because utilities manage critical infrastructures for social welfare and fairness. Climate change is likely to increase both the importance of resilience and the challenges inherent in providing a reliable electricity supply. One only needs to point to the recent 2023 Canadian and 2020 Californian wildfires, 2021 Texas winter storms, 2023 European and Chinese summer droughts, etc., for evidence of this.

To incentivize utilities to invest in climate resilience, one immediate step is to quantify the climate resilience of electric utilities (“utility resilience”). In this thesis, we argue a combination of regulatory benchmarks and machine learning (ML) will be pivotal [30]. In general, government regulation of utilities is necessary given they are often treated as natural monopolies, at least for some part of their production process. Regulation of utility resilience is complex due to the dual complexity of both power and climate systems, as well as the non-linear and high-dimensional dynamics between weather impacts and power outages. These challenges make ML an ideal method for regulatory benchmarking, as modern ML models are highly optimized to approximate non-linear relationships while digesting a multi-dimensional set of data through efficient algorithms. However, under the changing climate and our societal-wide decarbonization (especially electrification), we have limited regulations that explicitly target (vs. implicitly consider resilience as a co-benefit of grid modernization like in Hawaii [29]) utility’s resilience outcome (vs. identify resilience actions like in Italy [7]). Therefore, exploring the use of ML-based benchmarking to regulate utility resilience could also offer novel insights for future developments of regulatory tools.

The rest of thesis will be organized as follows: section 2 paves our methodological foundation to synthesize ML into regulatory benchmarks by reviewing and discussing how both ML and benchmarking have been applied in the regulatory domains; section 3 offers background to our specific case of utility resilience regulation and the potential implications of applying regulatory benchmarks via performance incentive mechanism

(PIMs); section 4 applies our methods into a utility resilience case through a set of experiments on benchmarking utility resilience across distribution network districts in New England between 2010-2021.

Chapter 2

Regulatory benchmarking by machine learning (ML)

This chapter describes the two pillars of our methodology, ML and benchmarking, in the context of government regulation. For each pillar, we start with a literature review on its existing status and then relate it to our application cases.

2.1 Regulation by ML

Before we apply ML to the narrow field of regulatory benchmarking, it is worth discussing how ML-based regulation works in general. How has ML been applied in the related literature? And, why should we explore regulating by ML in the climate resilience space in the first place?

2.1.1 Literature review

Table 1 below summarizes highlights of recent literature on ML-based regulation. These studies are mostly qualitative with the majority of them coming from authors or publication venues in the domains of law, political science, and public policy.

Previous studies can be divided into two categories: the internal ML-regulator dynamics and external implications of ML-based regulation. These categories can be further split into five fields, as outlined in table 1 and expanded in the following subsections. Internally, some papers put ML as a subject and investigate its impacts on regulators, while others focus on regulators and study how they use ML as an object. Externally, academics mostly focus on three aspects that stem from regulation by ML: law scholars' study of its constraints, policy analysts on its applications, and government agencies' initiatives.

Given that our thesis focuses on methodological applications of ML-based regulations, we further expand our reviews on the internal dynamics between ML and regulators.

	Theme	**Paper highlights**
Internal dynamics between ML and regulator	ML impacts regulator	• ML reduces labor and complements humans [17]. • ML makes official's rules more consistent and precise [6]. • ML's ethical challenges require government to first develop capacity, integrate models, and tackle inequalities [34].
	Regulator uses ML	• ML is like a new machine whose promises and dangers fully depend on government users [15]. • ML is a tool for government efficiency and it mirrors citizens' interaction with similar tools in civil affairs [35]. • ML are primarily models to exploit administrative data and improve daily operations rather than policy evidence [54]. • ML are generic tools for core governance functions, including detection, prediction, and decision-making [34].
External implications of regulation by ML	Constraints	• Regulation by ML has to be under the rule of law with adversarial interrogation integrated [24]. • Regulation by ML mostly falls under the administrative law rather than constitutional principles [18]. • Regulation by ML is constrained by its conflicts with core values of administrative law [6].
	Applications (in theory)	• Specific: environmental monitoring [25], clerkless public offices [60] • Generic: finance, healthcare, transportation [48], public service operations [59]
	Initiatives	• Administrative Conference of the United States [18] • European Commission's Joint Research Centre [36] • Estonia's Government Chief Technology Officer [53]

Table 1: Literature on machine learning-based regulation.

One stream of literature treats ML as a new technology that just like many technologies, brings pros and cons. In [17], ML is framed as a cornerstone of AI as a transformative cognitive technology. It emphasizes how ML could automate a diverse set of government operations through relieving (e.g., subway system planning), splitting up (e.g., chatbot-plus-human online Q&As), replacing (e.g., address extraction in mails), and augmenting (e.g., real-time CCTV camera monitoring) officials' work. It also quantifies the potential cost saving of optimistically 1.2 billion working hours or an equivalent \$41.1 billion per year. Moreover, ML creates a precise standard that can be applied consistently by agencies. However, such a standard brings issues of explainability which even the most experienced officials could not grasp its intuition. [34] proposes several remedies regulators must take when adopting ML, including enhancing in-house technical capacities, where agencies historically tend to outsource to external contractors. The study points out that to tackle future silo-like crises such as the COVID-19 pandemic, regulators have to sharpen their data pipeline and model integration as well as actively monitor structural inequalities that could be easily aggravated by ML models.

Another stream of papers views ML as a tool under the current regulatory framework and compares it to other types of regulatory tools. An early work by [35] focuses on the connection of ML with other tools that have been used by the government or are being used by citizens in their personal lives. The authors encourage the government to explore ML as a way to synchronize with the public on adopting emerging digital tools. Diving deeper, [15] sets the stage with a strong argument that ML is just a new machine. The alarms raised by activists against ML-dominated regulations should be perceived as reminders for deliberate and careful implementation rather than prohibition. In contrast, [54] and [34] take a more neutral stance on regulation by ML and focus on the new possibilities of functions brought by ML such as an updated version of e-Government and enhanced governance on rights protection.

2.1.2 Why not regulate by ML

One missing piece within the literature that we discussed above on ML-based regulation is the rationale of adopting ML for regulations, particularly from the perspectives of ML researchers. Here we attempt to briefly uncover the rationales from three dimensions where the paradigms of government regulation and ML systems might parallel: complexity, information, and optimality.

When regulators and ML systems make decisions, both types of agents have to face complexity, but in different forms. For regulators, their challenge lies in the complex sociotechnical system, while for ML, the complexity is unusually presented as large-scale and high-dimensional data. To tackle these complexities, both resort to some mode of simplification. For regulators, they are granted discretionary power within the complex space that cannot be further bounded by written rules [24]. For ML, researchers push for dimensionality reduction that reduces the number of dimensions or variables in the input data. On the other hand, their simplification processes are both bound by external rules, which is critical to yielding decisions that are compatible with the reality of complexity. For regulators, their discretion power corresponds to their obligation to explain their decisions when being called upon, which in return helps qualify or disqualify their discretionary competence [24]. Similarly, ML models, which could be represented with up to trillions of parameters (e.g., GPT-4) and theoretically have a near-infinite dimensionality reduction power, are still bound by their out-of-data generalization and computing costs.

Additionally, both regulations and ML heavily rely on information, i.e., evidence and data. For regulators, evidence is usually expected when regulatory decisions incur complaints or litigation, although not all regulatory decisions are based on the input of external evidence. For ML, the state-of-the-art models are not only internally driven by data but externally evaluated by data. For example, 5 out of 6 metrics used by the Stanford AI Index to score the ML model's technical performance are based on benchmark datasets including the ImageNet which consists of millions of images [61].

Although evidence and data are external, they are not necessarily objective. Indeed, evidence and data are political in both contexts of regulations and ML. For instance, during litigation, judges can down-/up- weight or even exclude/include evidence presented by either regulators or regulatees. And of course, regulators or regulatees can make similar choices when presenting their evidence. Not surprisingly, ML developers can also manipulate both input data (e.g., features, resolutions) and output samples (e.g., testing set) to influence external validity.

To legitimize regulation, [6] proposed it must reflect means-end rationality as the core value of administrative law. We argue that ML could serve as an "optimal" means to achieve regulatory ends. ML models or algorithms are mathematically optimized in a way that ML calculates parameters that could maximize (or minimize) some objective function such as maximum likelihood estimation (or mean squared error, MSE). Such an attribute of ML, (local) optimality is provable and logical. In contrast, regulatory proceedings can often be arguable and non-empirical. ML's local optimality could still be leveraged as a form of short-term rationality. In other words, instead of asking "How can we optimally predict the long-term future“, regulators could ask "How can our short-term optimal predictions inform our actions for a long-term optimal future?" This emphasizes the active role of the regulator, instead of ML, within the decision-making loop. And [15] showed that such a process, where regulators are actively rather than passively involved, could make many legal challenges untenable.

Like many other scenarios of ML applications and regulatory tool adoptions, ML is not a silver bullet for all regulations. As discussed by [17]'s quantitative analysis and illustrated in [60]'s imaginary clerkless offices for administrative affairs (i.e., an augmented Amazon Go store), ML can save regulator labor by automating daily office operations. In cases like making individual-level judgments [6], regulators' discretion is however crucial. When comparing the capabilities of ML with regulators', we should be cautious to avoid being trapped to the extreme ends of opinions, either anxieties or faiths, in ML [18].

ML researchers or developers who team up with regulators should transition from a data-driven to a data-literate mindset that understands both the data and its policy implications. Similarly, regulators should treat data not only as a tool but also as a science that requires basic training. For example, it is tempting for regulators to leverage their large data repository [54] to unleash ML power, but recently ML researchers have demonstrated that good data could sometimes be more important than big data in order to build a practical ML system [52]. Our data literacy should also alert us that all data are sampled in a specific context and their regulatory validity for ML should therefore be restricted within such context [42]. The effort to improve data literacy may need to go beyond talent education to talent attainment, particularly as today's public-private ML capability gap widens. As [43] referred to in 2018, an average "data scientist" position in the US pays $120,931 per year, and to match that salary in the federal government, this position has to rank at GS-15, the most senior rank of white-collar workers.

Lastly, there are nuances that ML and regulation practitioners need to revisit from the famous metaphors of the regulator as "graybox" and ML as "blackbox". Many ML models are indeed hard to intuitively explain, but technically, ML models' algorithmic reproducibility and numerical precision do make them highly transparent. In this case, ML has an advantage over the partially explainable processes of regulatory rulings, and should better be called "blackglass". Indeed, [19] suggests that when ML algorithms encode legal principles and agency priorities, they could even qualify as rules under the administrative law via notice-and-comment process or pre-enforcement judicial review. Sometimes explainability might yield to other regulatory priorities such as the efficiency of information analysis and costs of internal management, which indirectly contribute to regulatory decisions but do not require explaining their intuition. In these cases, we can treat both regulators and ML as plain "boxes" for internal decision-making uses, and divide their labors via cost-benefit analysis. For instance, the Securities and Exchange Commission replaced its staff with algorithmic tools to more efficiently scan accounting reports and trading data to flag potential securities law violators [19].

2.2 Regulatory benchmarking

2.2.1 Literature review

Studies on regulatory benchmarking emerged in the late eighties when information economics and agency theory were introduced into economic regulation [5], including the development of yardstick competition [50]. The theoretic foundations of regulatory benchmarking have since been well established in the domain of public utility regulations, with particularly widespread practice over electricity distribution operators in Western Europe [33].

One school of rationale on regulatory benchmarking focuses on its economic theories where benchmarking is used as a tool for incentive regulation [27]. Here, the utility's performance is compared to a benchmark's, and incentives to reach or exceed the benchmark are created. For example, a climate resilience benchmark, like the one developed below, could generate a benchmark based on the weather faced by a utility over some time period. One scenario is that the utility profits from having an actual resilience outcome better than the benchmark, and is fined if they underperform the benchmark.

An incentive benchmark could also be built over peer utilities' performance [5], or even a measure of the "frontier" of cost efficiency by only benchmarking over those top-performing peers. For example, European regulators have widely adopted frontier analysis methods, such as data envelopment analysis and stochastic frontier analysis, to quantify efficiency requirements in their price cap regulations [4]. Benchmark-based competition can introduce anti-competitive risks, as regulated firms can be incentivized to influence benchmarks through their definitions and selections of models, variables, etc. [28] In the case of German regulations on gas distribution operators, it was found that their benchmarking results were highly sensitive to how cost driver variables were selected and whether network usage was specified besides network expansion [55].

Another rationale for regulatory benchmarking is based on the principal-agent game between regulators and their regulated firms [5], where regulatory success hinges upon the alignment of information and incentives from both sides. Information asymmetry is a fundamental challenge in regulation, especially over internal corporate operations such as cost and efficiency, where the regulated firm has more and better information than their regulators. Regulatory techniques such as yardstick competition could alleviate this issue by developing a benchmark that infers a firm's attainable costs from cost data collected over a pool of similar firms [46], though regulators need to carefully tailor their collection strategies to avoid information overload [27]. Moreover, regulated firms like public utilities often serve as critical infrastructures whose regulatory governance is highly complex and dynamic. By constraining their regulations around benchmarks that target specific performance metrics, known as performance-based regulation (PBR), regulators could provide explicit and directed incentives to their regulatees [49]. We also need to strike a regulatory balance when aligning information and incentives at the same time. For instance, regulators and their consumer taxpayers may prefer utility services that are differentiated by local communities, but a more differentiated service profile makes its benchmark less effective in capturing the industry-average utility performance [8].

2.2.2 Integration into performance incentive mechanisms (PIM)

Given a benchmark on a regulatee's performance, how could regulators turn it into incentives to motivate performance? One way is to integrate the benchmark into PIM, a regulatory tool with particular applications in service qualities of electric utilities, and it also recently drew attention from state regulators in the US to regulate utility resilience [23]. PIM is a regulatory mechanism that financially rewards or penalizes a regulated firm, usually through its earnings or revenues, based on its measurable performance metric in an area of regulatory interest and with respect to a regulatory goal [49]. It is one of the four major approaches in the PBR toolkit [49] and it stands out for its explicit incentives, targeted improvement, and flexible implementations [58].

There are many ways to incorporate regulatory benchmarks into PIMs, and in figure 1 below, we illustrate the intuition of one way by borrowing the "yardstick" notion from the yardstick competition mechanism. The key difference is that yardstick competition uses equivalent firms' performances as "yardsticks" to mimic market competition [50], while our version of PIM uses benchmarks as "yardsticks" to help align, like in hurdling, a firm's actual performances with regulatory goals. That is, we can first build models as "yardsticks" to approximate, not perfectly measure though, a performance benchmark based on a firm's historical performances. Then regulators could apply their discretion [30] to adjust the stringency of aligning their final regulatory goal with respect to the initial "yardstick" — say being lenient in the first year to only require the firm to reach 60% of "yardstick" while pushing for incremental improvement towards 70% of the "yardstick" next year. Depending on how far the firm over- or under-performs the regulatory goal, they will be awarded or penalized proportionally. Lastly, the "yardsticks" can be dynamically updated (e.g., by retraining an ML benchmark model over additional data from recent years) to reflect the firm's changing performance potential.

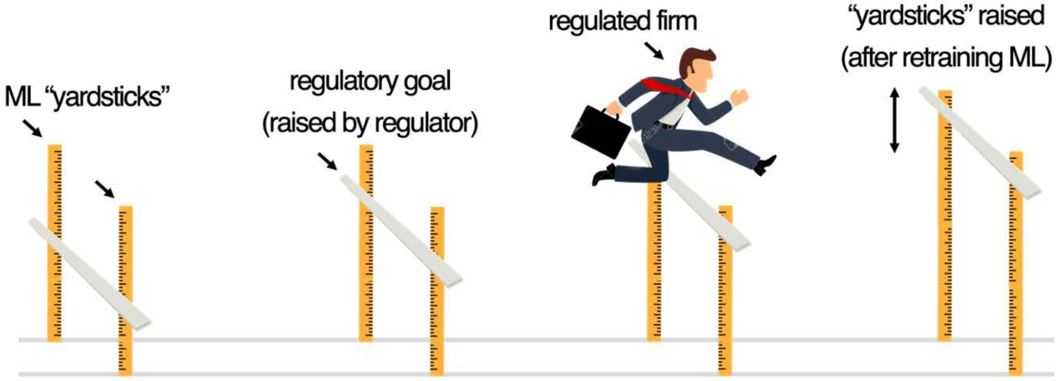


Figure 1: Illustration on performance incentive mechanisms.
Incentivizing a firm's performance with regulatory yardsticks (e.g., machine learning (ML)-based benchmark models) — like hurdling.

Chapter 3

Regulating climate resilience of electric utilities ("utility resilience")

3.1 Utility resilience

Here we define climate resilience as a system's ability to adapt to, withstand, and grow from impacts caused by the climate system (shortened as "climate impacts"). Resilience emphasizes the temporal aspects of system responses, especially its long-term improvements before, during, and after climate impacts. It is also a systems-level concept that implies the complexity of pathways toward its mitigation [16]. For climate impact, we broadly define it as any impacts that stem from the climate systems, and at the level of utility resilience that this thesis is scoped on, exhibit as local weather impacts. Climate change is increasingly driving today's climate impacts and the societal impact of climate change largely acts through weather impacts [3].

In terms of electric utilities (shortened as "utilities"), they are complex systems that possess different organizational and technical structures, and they continue to evolve. Currently, there are around 3,300 utilities in the US and most of them are privately investor-owned, while the rest are either publicly owned by local governments or privately owned by small-scale consumer co-ops [49]. Investor-owned utilities are this thesis' focus, not only because they serve about three-quarters of the US population [49], but because their non-public and non-consumer ownership design obligates them to be heavily regulated by state and/or federal government agencies, which will be detailed in section 3.2. On the technical side, a utility system generally consists of four major modules: electricity generation, power transmission, distribution networks, and consumer demand [1], though US utilities' generation function has been largely displaced during the 1990s "electricity restructuring" [9]. Each module is facing its own changes. For example, the booming technical advances and cost reductions of renewable energies and energy storage technologies are replacing base generation sources such as natural gas and nuclear plants very quickly, while new digital technologies are reshaping a smart grid throughout all modules and informing consumer behavioral changes from passive consumption to active participation [49].

The dual complexity of climate impacts and utility systems is making utility resilience management highly challenging, which is further worsened by the continuing climate change. Here are some examples before we will expand on their economic implications in subsection 3.2.2. The increasing prevalence of drought is draining reservoirs for hydropower generation, a critical source of power generation especially in developing countries [56]. Similarly, the increasing intensity of heat waves is reducing the transmission capacity of power lines and their secondary effects on wildfire enhancement are also threatening the distribution networks [1]. More recently in February 2021, the severe low temperature in Texas led to an unprecedented demand surge and pressure on the un-winterized grid, which further triggered large-scale electricity cutoffs [11]. These individual observations of climate impacts are not coincident, and they have been intensifying in recent decades, as evidently shown by the increasing frequency and intensity of Atlantic hurricanes and areas of wildfire burning in the US [10]. The long-term future of climate change continues to show worrying signs as well. For instance, rising temperature could reduce the average summertime capacity of transmission lines in the US in 2050 by nearly 5.8% as compared to the 1990-2010 period, while countries that heavily rely on hydropower such as Congo and Zambezi may see a decline in their hydropower generation by as much as 6.5% by the end of this century even under the assumed compliance of Paris Agreement [3].

3.2 Utility regulation

3.2.1 Current regulation in the US

Utilities usually warrant government regulation both economically and politically. On the economic side, their investments tend to require high upfront costs (e.g., a substation) and their operations exhibit network effects (e.g., the more customers on a distribution network, the more valuable it is) (D. Hsu, personal communication, Feb 2022). These characteristics make decentralized competition between utilities economically inefficient compared to a market with centralized and vertically integrated utilities. This structure increases the market powers of utilities, requiring regulatory interventions. Socially, utilities serve as an infrastructure that underpins almost every aspect of residential living, business operation, and industrial production, so their operational efficiency and security are politically sensitive and motivate regulatory actions from policymakers.

In the US, current utility regulation is highly heterogeneous in terms of its regulatory agencies, regulatory targets, and regulatory mechanisms. There are three clusters of agencies, including the federal-level Department of Energy (DOE) on technical development and advising, the Federal Energy Regulatory Commission (FERC) (and its semi-governmental subordinate, the Independent System Operator) regulating crossstate utilities, especially on standards and transmission [1], and state-level public utility commissions (PUCs) regulating within-state utilities, especially on distribution, demand management, and the building of generation (when the utility owns generation). There are also two major regulatory targets — the rate of electricity pricing and the quality of electricity service (especially interruptions) — that mostly interest these agencies. The corresponding regulatory mechanisms will vary depending on the regulatory target. The two mainstream mechanisms, and this thesis' focus, are cost-of-service regulation (COSR) mostly on rate, and PBR mostly on service quality, though there are a few other less-used regulatory tools such as integrated planning based on risk evaluation, permitting new lines of utility services, and tariffs to leverage private investment [29].

COSR is the classic and dominant regulation for US utilities. It allows utilities to petition the PUC to adjust the rate they charge their electricity consumers, utilities' major source of revenue, to fully cover their service costs along with a profit margin [49]. COSR is advantageous when there is a strong growth of electricity demand, as utilities can expand to meet these demands at minimal risks given highly certain payback from consumers through rate adjustments.

PBR has been defined in various ways, and here we define PBR as a regulation that incentivizes the regulated entity to achieve a specific performance goal without the regulator enforcing its achievement path. Several regulatory tools could satisfy our definition (see examples in [49]), but since this thesis narrowly focuses on utility resilience, we base our discussions on a PBR tool, PIM (see its definition and intuition in subsection 2.2.2), that is most classic yet well illustrates our following discourse on utility resilience regulation. This PIM tool consists of three core quantitative components: a regulatory goal, a performance metric, and an incentive design to fill the metric-goal gap [58]. The regulatory goal relates to some aspects of utility performances such as grid reliability and carbon emissions. The performance metric could be directly measured or indirectly inferred or modeled. The design of incentives is highly flexible, though it usually includes a rate of financial reward or penalty, and possibly, a rate of discretional adjustment on the metric [30]. See an illustration in subsection 2.2.2 for intuition.

Besides PBR's compatibility with regulatory benchmarks, PBR (particularly PIM) interests us for several additional reasons. First, utility regulation has been historically dominated by the COSR-type frameworks and after its debut in the 1980s [44], PIM offers a brand new "technology" of utility regulation that is worth experimenting with. The welcoming sentiment is reflected in today's surging rate of PBR adoption across at least 17 US states [49], especially on state-level distribution- and demand-side regulation [23], as well as the 2020 survey of major US utilities, where about one-third of respondents favor more PBRs or PBR-hybrid regulation [41]. Second, the PBR enthusiasm reflects the changing priority of utility regulation, which is shifting away from utility expansion (and the use of COSR) given the flattening electricity demand over the past decade [49], the fundamental rationale of COSR. Instead, utility regulators are facing a more diverse yet complex set of goals including grid decarbonization [31], energy justice [21], and climate resilience, as well as their confluence with technical upgrades aforementioned in section 3.1. Lastly, PIM is analogous to another emerging technology, ML, since both of them are goal-driven (i.e., a PIM performance metric vs. an ML objective function) while affording significant flexibility in the means (i.e., PIM-regulated firm's discretion vs. ML-ed "curve-fitting" blackbox models) to achieve their goals.

3.2.2 Why regulate utility resilience

Now that we have defined utility's resilience and its regulation, we naturally ask: why do we need to regulate utility resilience?

One core argument is the reduction of outage costs from utility resilience regulation. Weather is the major cause of power outages, responsible for 75% of outage durations in the US [?]. If we can strengthen our utility resilience, we can reduce weather-caused outages and thus the majority of outage costs.

Another key argument is based on the large potential impacts on the economy and health from a lack of resilience. When an outage happens, it introduces costs that are both economic (e.g., factory shutdown) and non-economic (e.g., air conditioning off). These costs are particularly salient during long-duration widespread power outages, which are almost all caused by extreme weather, and their external costs can be easily overlooked in decision-making [32]. Additionally, these costs raise questions of fairness given the growing evidence that shows the unequal resilience between communities of different demographics (e.g., the 2021 Texas winter outages on people-of-color neighborhoods [12]).

The third reason is about climate risks. The reliability standards in the public utility law imply the requirement of utilities to improve climate resilience, while the foreseeability principle in the tort law also obligates utilities to reduce avoidable harms of their services, including outages caused by the well-studied anthropogenic climate change [57]. The worsening trend of climate change and its impacts on weather and power outages is significant, yet highly uncertain depending on our decarbonization trajectory. The risks of evading resilience planning are therefore very high and justify more government regulation (e.g., the US government's chief in-house watchdog, Government Accountability Office (GAO), started lamenting DOE and FERC for their inaction of regulatory oversight in 2021 [1]).

3.3 Benchmark-based PIM for utility resilience

3.3.1 Advantages

Below we only list advantages that stem from the specific use of PIM on utility resilience, rather than those that can arise from the use of any generic regulation, as they have already been enumerated in subsection 3.2.2.

3.3.1.1 Informed and efficient improvement of resilience

To mitigate risks like climate impacts on utilities, the first step is usually information sharing on our progress, which is exactly PIM's strength. Since the regulatory goal and current performance metric are explicitly formulated in the PIM incentives, regulators can directly track where we are and how far are we from mitigating these risks. As compared to means-based regulation such as COSR and command-and-control on specific resilience investments whose information is partially accessible, these types of PIM progress information are highly relevant to the general public (e.g., System Average Interruption Duration Index, SAIDI, could tell us "how many minutes of outages I had this year") and could help quantitatively align the interests of the regulator, utilities, and the general public. Moreover, PIM's transparent metric, goal, and the metric-goal gap can enable many other forms of risk shielding options, such as informing individuals to freely decide to move away from under-resilient communities. The simple and transparent design of PIM also allows various stakeholders to experiment with it using different weather inputs, reward rates, stringency rates, etc. to explore counterfactual utility performances in different scenarios, which we will demonstrate through an online benchmarking portal in figure 3.

Furthermore, as compared to means-based regulation like COSR, goal-oriented PIM is more likely to maximize our resilience efficiency. There has been some theoretical justification of PUCs applying COSR to indirectly regulate utility resilience through the exploitation of regulators' discretion on rate approval [37]. However, for classic principal-agent pairs like regulator-utilities, one challenge is the enduring informational gap between regulators and utilities (D. Hsu, personal communication, June 2022) — how could regulators know better than the utilities which resilience action is more efficient? Besides the informational gap, there also lies an interest gap between regulators and utilities, which is actually widened by COSR as it incentivizes utilities to maximize service costs, not the regulator's interests in resilience. PIM's performance design can also be tailored to various resolutions and dimensions to regulate resilience equity, reducing social costs for efficiency. PIM can easily zoom in by targeting its accounting scope over a specific region or community without changing our designs of metrics and rates. Similarly, we could change the time horizon of our PIM from yearly SAIDI to say 5-year SAIDI, which could better inform resilience investments given that many utility resilience projects have a high lead time and a corresponding delay in observing their ultimate effects.

3.3.1.2 Reduction of regulatory cost

Besides improving performance, regulatory cost reduction is another selling point of PIM. Given the dual complexity of climate and power systems as illustrated in section 3.1, there are multiple dimensions of weather inputs (e.g., resolution, coverage, variables) as well as their nuances (e.g., snow that falls around vs. well below 0 Celsius on power line icing) that utilities need to consider in their resilience investment decisions. Also, the links between weather and utility performance are highly nonlinear (e.g., wind gust's effect on falling trees over power lines), which adds extra complexity for the utility to evaluate their investment decisions. Now, if we use the traditional COSR, these complexities will not only intensify a utility's filings for rate adjustments (as found in many other similar applications of COSR where utilities have to make complex investment decisions [49]) but overload regulators with frequent and intense evaluations of those rate adjustment requests. In contrast, under PIM, regulators can better focus their limited resources on assessing the ultimate performance of utility resilience, not their means of achievement. Indeed, resilience incentives could be easily integrated into current service-quality PIMs (e.g., at Massachusetts' Department of Public Utilities [38]) by simply updating the regulatory goal with weather adjustments, which further reduces regulator's costs of devising a new set of regulations for resilience.

3.3.1.3 Reduction of service cost

PIM will not only reduce regulatory costs but also very likely utilities' service ones. Under COSR, the containment mechanism on utility's service costs is very weak [49] or even reversed given that utilities' profits are proportionally tied to their service costs through case adjustments, and thus the new regulatory priority on resilience could easily add costs. In comparison, under PIM, utilities are incentivized to reduce their service costs given that their financial returns are directly tied (assuming we decouple PIM from COSR on utility resilience investments and operations) to their ultimate resilience performance. Compared to regulator principals, firms as agents also have better information and stronger incentives to explore and exploit their lowest-cost pathways toward resilience, which PIM could easily facilitate. Moreover, by explicitly clarifying regulatory priorities in PIM [49], utilities could streamline their organizational processes towards these priorities with reduced management costs.

With PIM taking effect, it might reduce the utility's service costs by creating an internal "market" within the utility to "trade" resilience [14]. Suppose the regulator targets a certain total amount of weather-adjusted power outages this year on a utility and this utility serves two districts (i.e., A and B) with an equal number of customers but significantly unequal costs of restoring from weather outages (e.g., A costs more than B). Now under PIM, the utilities could meet its performance goal with a lower cost by simply trading outages between A and B. Furthermore, if district A's customers' willingness to pay (WTP) for weather outages is lower than district B's, then the utility could even incur the same amount of outages in both districts as long as their sum meets their resilience goal while charging customers at district B a higher rate for their extra WTP on resilience than those at district A [16].

While these internal tradings enabled by PIM could reduce utility's service costs, they might introduce unanticipated social costs of fairness. District A is truly more costly to restore weather outages but this might be caused by the historical underinvestment of utilities over this district's utility infrastructure. In this case, the internal trading mechanism actually creates systematic discrimination against district A's resilience in the long run. On the other hand, it might be also true that customers at district A seem to value less on weather outage's damages and climate risks given their lower WTP, but is this value difference caused by their intrinsically different assessments, or externally different market opportunities to pursue jobs that require more resilience? Just like Viscusi's dilemma on values of life [40], these types of opportunity-fairness vs. cost-efficiency tradeoffs deserve careful deliberation (e.g., by adding a minimum standard of resilience across districts) before we apply PIM for utility resilience.

3.3.1.4 Innovation for resilience technologies

Given PIM's demand certainty and methodological flexibility, it sets up a giant and liquid market that could cultivate innovation in resilience technologies. Previous applications of PIM on generic service quality have promoted new electronic systems such as the Fault Location, Isolation and Service Restoration (FLISR) which could isolate faulty distribution networks and transfer their electricity loads to other healthy networks so that the number of customers interrupted by outages could be minimized [51]. Similarly, given the high risks in hurricane-prone areas such as Florida, utilities could experiment with risky capital projects such as undergrounding their power lines [16] given the discretion of PIM on utility investments.

3.3.2 Potential Pitfalls

Even though we believe the benefits of a properly designed PIM to regulate utility resilience significantly exceed their costs and risks, there are several potential pitfalls that we believe should be carefully examined and prepared for before enforcing PIMs.

3.3.2.1 Legitimacy of resilience goal, metric, and incentive

Regulatory benchmarks are promising given their highly quantifiable resilience goal, metric, and incentives, but what about the political economy behind them? Getting the political economy right is critical to ensure the legitimacy of our PIM framework, and here are several example issues we should consider. First of all, who decides our regulatory goal? As explained in subsection 3.2.1, utilities are cross-regulated by multiple federal and state agencies, each with unique jurisdictional responsibilities, political interests, and bureaucratic culture, so further studies are needed on how to coordinate these agencies more smoothly (e.g., weather could trigger power outages through interstate transmission lines or local distribution networks, with each regulated by FERC and PUCs respectively). Now, how do we adjust our goal based on weather? One way is to build an expert assessment panel of meteorologists and utility researchers, but the panel-building process is not free from politics. Another way is to build a computer model but depending on whether this model is correlative or causal, the interpretation of the model's prediction will be highly contentious, e.g., weather-related vs. weather-caused, significance of causal or correlative effects, etc. Lastly, the two rates of financial incentive and stringency are non-trivial, as these input parameters are susceptible to value judgments and could have multiplicative effects on the final financial reward or penalty.

3.3.2.2 "Overfitting" utility resilience

While applying PIM to utility resilience, we implicitly transfer climate risks to utilities through an incentive mechanism that centers around the utility's performance, rather than customer demands. Interestingly, PIM transfers risks in the opposite direction of COSR's, where utilities transfer their risks to their customers through rate adjustments. Both schemes of risk sharing are highly unbalanced, which could lead to moral hazards of consumers overloading the grid without activating their low-cost resilience preparedness (e.g., backup electricity generator [49]), or governmental agencies undermotivated to assist utilities under extreme weather [16].

Furthermore, PIM's nudge on utility resilience could spark other unexpected market and political failures. Given the PIM's transparency on utility resilience performance, as discussed in subsections 3.3.1.1 and 3.3.1.2, they can however divert attention and resources of various stakeholders away from other regulatory priorities such as grid decarbonization, the root cause of climate risks, as well as create a risky societal belief that PIM-regulated utility resilience can be used as an "insurance" against our slow decarbonization progress. We also need to be cautious about the "ratchet effect" [45], where utilities might be under-incentivized for performance improvement given that over-improving this year will likely raise regulatory expectations of their next year's improvement. This pitfall might be mitigated by adding regulatory goals relative to utilities' external benchmarks (e.g., peer utilities' current resilience), rather than purely internal ones (e.g., the same utility's historical resilience).

3.3.2.3 Regulatory uncertainty under uncertain climate

The climate system is inherently uncertain, which sharply contrasts with PIM's regulatory certainty, the cornerstone of the utility's long-term commitment to resilience investment [44]. One key input in the assessment of utility resilience is weather data, which can be roughly categorized into two types — point-scale weather station data and simulated grid-scale data. Each has its intrinsic uncertainties (e.g., the spatial sparsity of weather station data and the bias of simulated weather data). Additionally, weather patterns under the future climate are highly uncertain, particularly depending on our decarbonization progress. How could we incorporate these uncertainties into our PIM formulas? E.g., a confidence interval on our weather-adjusted regulatory goal.

3.3.2.4 Limits under extreme weather

The intensifying extreme weather events under climate change might test how far current PIMs could be applied. Extreme weather events are low-likelihood high-impact, which makes it particularly challenging to plan the grid for ex-ante, as well as attribute outages to extreme weather ex-post, both empirically and statistically. Moreover, extreme weather events are highly personal and perceivable, which could easily induce behavioral irrationality for various stakeholders [16]. In the following experiment section, we will illustrate this challenge of benchmarking under extreme weather again from an empirical lens.

Chapter 4

Experiment: ML-based regulatory benchmarking for utility resilience

4.1 Expectations

In this experiment, we aim to build an ML-based model that benchmarks utility resilience in New England. Due to constraints on time and computing resources, we focus on three network distribution districts (shortened as “districts”) within the operational territory of National Grid, a major utility in this region, but our approaches should be easily scalable to other districts. For the purpose of experiential learning, these three districts are anonymized as a, b, and c, respectively. And our benchmark specifically measures per-district daily resilience with respect to an average day in a historical period within the same district. There are other types of benchmarking goals with nuanced meanings, such as benchmarking against an average district-day instead of an average day, or even conditioned on the same weather, but their model implementations and regulatory implications would be drastically different.

Here we define utility resilience as weather-related System Average Interruption Frequency Index (SAIFI, $= CI/CS$, where CI and CS refer to customers interrupted and customers served respectively), with SAIFI quantifying the average number of outage interruptions a customer experiences within some time and space. There are several other similarly formulated reliability indices, including SAIDI ($= CMI/CS$, where CMI indicates customer minutes interrupted), which we didn’t use given its additional temporal component that might require minute-level weather data to capture.

Rather than building “another application of fitting ML models”, we impose several additional requirements on our benchmarks to make them truly useful for regulatory purposes. That is, a benchmark model that is accurate, accessible, and applicable. These three criteria were respectively inspired by Clark’s credibility, legitimacy, and saliency principles on environmental assessments [13], as well as the advocacy by Selin to use them to guide modeling for policy [47]. Our proposed criteria are more tailored to the side of modeling practices.

An accurate benchmark model should be ideally unbiased within its reference (e.g., training set's) distribution, with bias defined as the average error term, i.e., $\bar{\epsilon} = \bar{y} - \bar{\hat{y}}$ [30], where y and $\hat{y}$ indicate groundtruth and prediction respectively. This property can be satisfied by any, but not exclusively, model that has an intercept term and is least-squares optimized. These conditions do, however, restrict us to analytic or numeric, not algorithmic, models. Just like any predictive ML model, an accurate one should fit its groundtruth well, which can be measured by the coefficient of determination (R2), amongst other measures. In addition, a benchmark model should not just be internally accurate with regard to its given data, but externally authentic by basing on high-fidelity data input in the first place.

We also expect the benchmark model to be highly accessible throughout its development stack, particularly data access, feature engineering (if needed), and model parameterization. Open-source and free-access datasets will be prioritized. Models that are parsimoniously parameterized and highly interpretable are preferred, and this rule applies to feature engineering as well. These traits are particularly important for the benchmark to be adopted by regulators and utilities, whose literacy and resources on ML usage are likely constrained.

Lastly, a useful benchmark model should be able to be practically applied to real-world regulations. To maintain regulatory certainty, the benchmark model's output should have minimal anomalies such as rates of negative predictions on SAIFI, which in theory should be non-negative. The valid range of outputs should also be wide, especially with a high upper cap that can capture high SAIFI days under extreme weather. To maximize both research impacts and regulatory adoptions, we also hope to build downstream demonstrations on how our benchmark models could be deployed at a large scale.

4.2 Data and features

All relevant datasets that have been used in our experiment are specified below in table 2, where we summarize their type, spatial/temporal resolutions/scopes, as well their corresponding features we can derive.

The Local Climatological Data (LCD) weather [22] and outage data products ended up being used in our final models. Based on outage event data (i.e., CMI, CI) at the point-minute scale, National Grid assisted us by aggregating them into the district-daily scale along with corresponding metadata (e.g., CS). From there, we calculate daily per-district SAIFI as our target variable and further normalize it by dividing

Data product	Type	Spatial		Temporal		Derived features
		Resolution	(Our) Scope	Resolution	(Our) Scope	
Outage (from National Grid)	Groundtruth outage events, including customers interrupted (CI), customers minutes interrupted (CMI)	point (by transformer)	NY, MA, RI	<= minute	2010-2021 (2012-2021 for ENTLN)	Daily per-cell (e.g., 15 x 15-km grids) outage summaries (e.g., SAIFI, SAIDI), Daily per-district (of distribution network) outage summaries
Earth Networks Total Lightning Data (ENTLN) [26]	Groundtruth lightning events, including intra-cloud and cloud-ground flashes and their peak currents	point		<= millisecond		Daily per-cell lightning summaries (e.g., number of flashes, maximum current/ duration), daily per-district lightning summaries
Local Climatological Data (LCD) [22]	Groundtruth weather station observations (e.g., wind speed/gust, temperature, precipitation) by stations (mostly airports)	point (by station, ~4 stations per district)		sub-hourly		Daily per-station weather summaries (e.g., dry/wet snow, rain, top-3 hourly precipitation, number of wind gusts >30 or 40, temperature gradient)
University of Idaho Gridded Surface Meteorological Dataset (gridMET) [2]	Modeled weather (e.g., maximum/minimum temperature, precipitation, but no maximum wind speed/gust) as a grid	~4 km		daily		Daily per-cell weather aggregations (e.g., from 4-km to 15-km)

Table 2: Data specifications.

it by its standard deviation. For the LCD weather data, we first standardize them at the hourly per-station resolution by grouping from neighbor stations (i.e., shared latitude-longitude by 2 decimals) and sub-hours. Then we quality-control stations by removing those whose recorded numbers of hours or days, especially on maximum sustained wind, are sparse. We further aggregate the LCD data from hourly to daily resolution, where we differentiate daily precipitation by their physical states as rain, wet snow, and dry snow by summing hourly liquid-content precipitation at the temperature cutoffs of 29 F and 35 F. We also consider the intensity of precipitation by averaging the top three hourly precipitations. Similarly, we incorporate the intensity and duration of wind gusts as numbers of hours when wind gusts are above 30 mph and 40 mph. The changing curve of hourly temperatures throughout the day is also quantified using the temperature gradient from the least-squares fitted linear regression (nullified on days with less than 20 hours of recorded temperatures). For features that are still null, we fill them with the medians of nearby stations' same-day features. One last weather feature we add is the maximum sustained wind speed, which is out-of-box aggregated at the daily resolution from LCD.

There are several other types or products of datasets that we did not adopt but did inform us on finalizing our data pipeline. In our pilot studies, we explored using the Earth Networks Total Lightning Data (ENTLN) [26] as part of our weather features, but when conditioning on other weather variables such as wind and precipitation, these lightning features turned out to be not very significant in terms of both their coefficients and improvements on predictive performances. In our early stage of experiments, we also examined multiple grid-based weather datasets, particularly the University of Idaho Gridded Surface Meteorological Dataset (gridMET) [2], before trying the station-based ones. These grid-based datasets had to be excluded due to their low-fidelity observations on the wind, as well as their lack of hourly trends of temperature and precipitation.

4.3 Methods

We adopt ridge regression (RR) as our final model. It is a classic variation of linear regression with L2 regularization on its least-squares objective function, MSE, i.e., $\|y - Xw - b\|^2 + \alpha * \|w\|^2$, where $X \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}$, $w \in \mathbb{R}^n$, and α are our features, intercept, coefficients, and regularization hyperparameter respectively, for m examples of district-days and each with n features. In our case, we have $n = 768$ features and they consist of the 8 vanilla features per each of the 4 nearest weather stations (to incorporate spatial heterogeneity of weather inputs) we match with each district, their squares (to capture second-order weather effects), per-station wind vs. precipitation vs. temperature interactions, and their quarterly adjustments (to indicate seasonality), as well as adjustments on sustained-wind-related terms at the cutoff of 30 mph (to capture nonlinear wind effects). Each of these features (except dummies) is separately z-score normalized by subtracting by its mean and then dividing by its standard deviation.

We also customize a workflow to train, validate, and test our RR model. For each district's dataset, we use 2010-2016 for training-validation while 2017-2021 for testing. For the training-validation set, we further divide it into five folds through a prorated splitting at SAIFI cutoffs of $[0.01, 0.05, 0.1]$ so that days at each range of cut could be evenly distributed across folds, which is critical to improve our model's performance as we approach the tail end of SAIFI. Within each cut, we assign days greedily to folds of the least sum of SAIFIs, which additionally ensures a uniform sum of SAIFs per fold. To train our RR model by district (via a Python package, Sklearn), we tune for the optimal hyperparameter α by choosing the one that yields the lowest average MSE from each of the five validation folds when being trained over the rest of four-folds. The model trained with the optimal α on the whole 2010-2016 training-validation years is then tested over 2017-2021 (i.e., "outside-sample testing"). Given the limited number of days in our testing years and the potential distribution shift of resilience from 2010-2016 to 2017-2021, we further conduct a "within-sample testing" on the training-validation years by holding out one day at a time as testing day and retrain a model on the rest of days to test on the hold-out day.

There are also several other model variants and their setups that we have explored. To capture the non-linearity and high-dimensionality of weather-SAIFI dynamics, we tested directly feeding raw weather features (including both grid- and station-based data) into neural networks to let the networks learn features, instead of feeding hand-engineered features into linear regressions. The former plan was discarded due to both its relatively low accessibility (e.g., parameter interpretation) and low applicability (e.g., hyperparameters tuning, consistency across extreme weather events) compared to RR's, even when both are similarly accurate. Additionally, we tried different resolutions of "parallel training" (due to the low cost of training instances of linear models like RR) such as training separately by each weather grid cell, each quarter, or even each month. Even though training at a higher resolution implicitly injects more spatial-temporal information to facilitate model learning, it reduces the amount of historical weather-outage events in the training set, especially on extreme weather events. We also carefully adjusted the spatial scope of our target variable, SAIFI, in our model design. Directly predicting state-level SAIFI is too ambitious due to the spatial heterogeneity of weather impacts and outage occurrences, while finely predicting cell-level SAIFI is too ambiguous due to the spatial sparsities of high-fidelity weather grid data and high-magnitude outage events. Instead, we predict SAIFI at an intermediate scope of district-level, and then if needed, aggregate district-level results into state-levels for regulatory accounting.

To demonstrate how our model could be scalably deployed, we built an online portal for different modes of benchmarking usage, as shown in figure 3. The web interface is designed using the Gradio package while the web server is hosted on HuggingFace Spaces. In comparison to local-served software, a cloud-based website is advantageous for its centralized and controlled running environment, as well as its low barrier of usage for any users who can access a web browser. To facilitate user interactions without requiring users to compile code and format data by themselves, we contain our ML benchmark as a singular app with the minimally necessary Input/Output specifications. More details will be discussed in the next section.

4.4 Results and discussions

Table 3 below summarizes our model's performances across different districts and years. Overall, our model shows high goodness-of-fit with R2s for both within-sample testing and outside-sampling years around 0.3, except for districts b and c in within-sample testing years where R2s are around 0.1. For models of all districts within- and outside-sample, their rates of negative predictions are also very low, ranging from 3% to 0%. Given the long-tailed distribution of SAIFI and weather impacts, which will be explained in figure 2, we have to cap the range that our SAIFI predictions could be applied at 0.2, 0.1, and 0.1 respectively for each of the three districts during both within- and outside-sample testings. These caps do introduce a trivial amount of bias (-0.0004, -0.0011, -0.0009 SAIFI/day for district a, b, c) during testing because our models are originally benchmarked over a training set with no caps (i.e., all days are used for training).

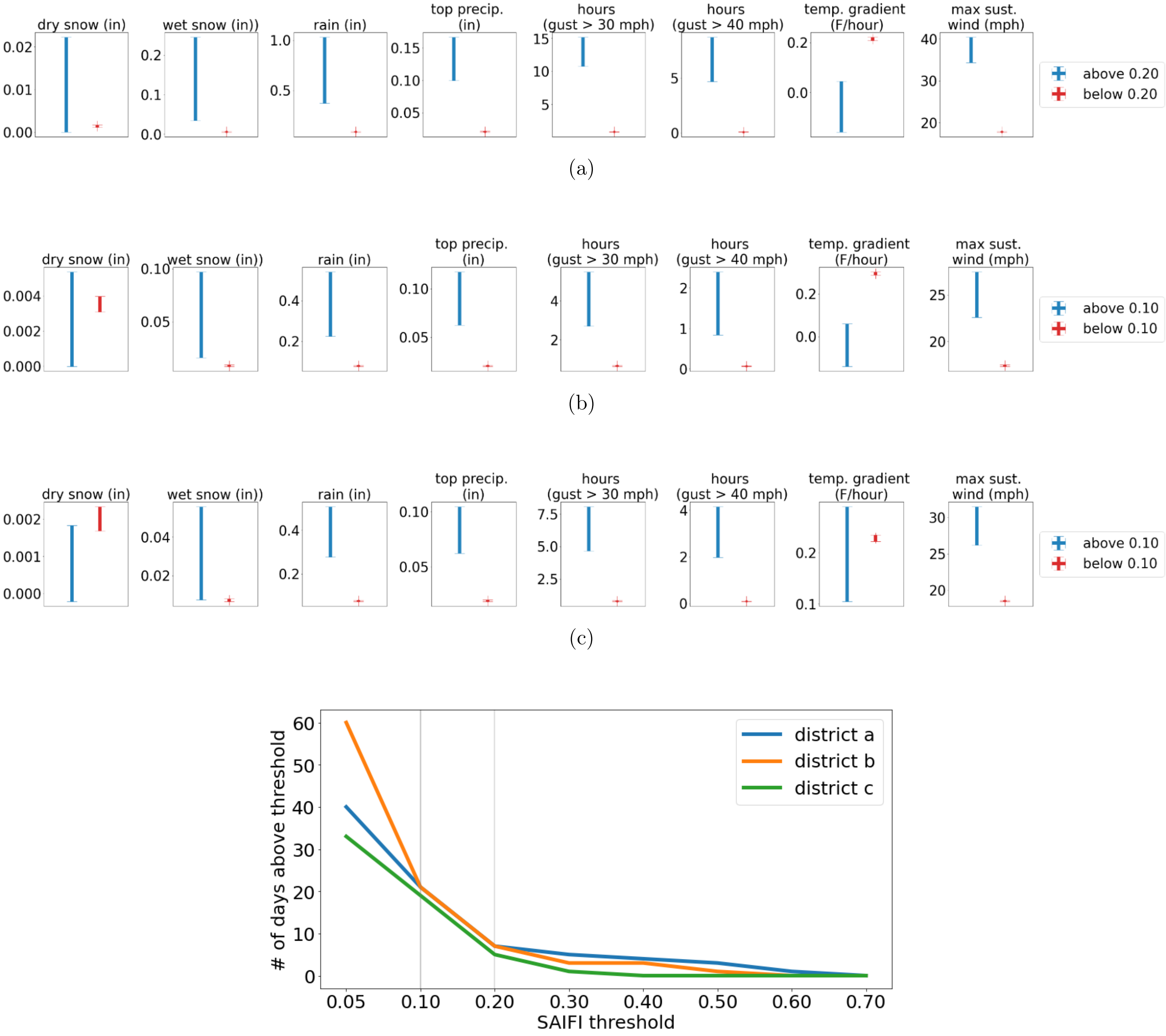
District	**Training (2010-2016)**				**“Outside-sample” testing (2017-2021)**		
		“within-sample” testing (leave-1-day-out)					
	Cap	Coefficient of determination (R2) (without cap)	Negative prediction rate	Bias (SAIFI/day) (after applying the testing cap)	Cap	R2 (without cap)	Negative prediction rate
a	None	0.383	2.9%	-0.0004	0.2	0.281	2.8%
b	None	0.106	0.6%	-0.0011	0.1	0.344	0.4%
c	None	0.068	0.0%	-0.0009	0.1	0.285	0.0%

Table 3: Model performances.
Model performances on benchmarking weather-related outages (i.e., System Average Interruption Frequency Index, SAIFI) across distribution network districts.

One bottleneck to improving our model's performance is the long-tailed distribution of both our outages and their weather conditions. As shown in figure 2.d, SAIFIs in all three districts are highly right-skewed, with the number of daily SAIFIs plunging under 10 after passing the 0.2 threshold. Furthermore, post the 0.2 threshold, the numbers of available daily SAIFIs at districts b and c approach zero at a faster rate than district a. This explains why we have to cap district a at a slightly higher magnitude (i.e., 0.2) than districts b and c's (i.e., 0.1), as weather and outage data on these high-SAIFI days tend to be highly informative on weather-outage correlations, despite their rarity. The lack of high-SAIFI days at districts b and c might also explain the lower within-sample R2s (i.e., -0.1) at district b and c, in contrast to district a's (i.e., 0.3).

Weather impacts are highly skewed as well. One way to parse out the skewness is to compare the mean values of 8 root weather features (across all stations per district) along with their 95% confidence intervals (i.e., ± 2 times standard errors) across districts a (figure 2.a), b (figure 2.b), and c (figure 2.c). We can see a significant disparity of nearly all weather features (except wet snow and temperature gradient for district c, and dry snow for all districts) between days below and above their assigned SAIFI caps.

So how to improve our model's performance on the tail? A recent survey [62] on improving visual recognition performances on long-tail distributed data outlines three major routes: class re-balancing, information augmentation, and module improvement. Even though that survey focuses on deep learning-based techniques, we can borrow its outline to brainstorm a few simple solutions in the context of regulatory benchmarking. First, we can naively up-weight the tail examples (and/or down-weight the head ones) to make our loss function more balanced during training so that our model pays more attention to learning outage-weather dynamics on the tail, but now, how to interpret the bias of our retrained SAIFI benchmarks as a sum? One simple trick is to logarithm-transform our SAIFI target varible, which will then change our interpretation of bias from its sum to the order of magnitude. Second, we can augment the information on the tail by transfer-learning from the observation-rich head, but how feasible is this given the prevalent head-tail weather disparities we have shown in Figures 2.a - 2.c? Third, we can improve our model design by decoupling it into two stages — representation (e.g., feature extractor) and classification (e.g., head/tail classification). We can further enhance our feature extraction by using an ensemble of RRs (e.g., [20]) across SAIFI categories, and it is straightforward to show that our final SAIFI prediction would still be unbiased if we weight (conditioned on being normalized across categories) the least-squares loss functions of each RR using each example's probability mass function.



(a)

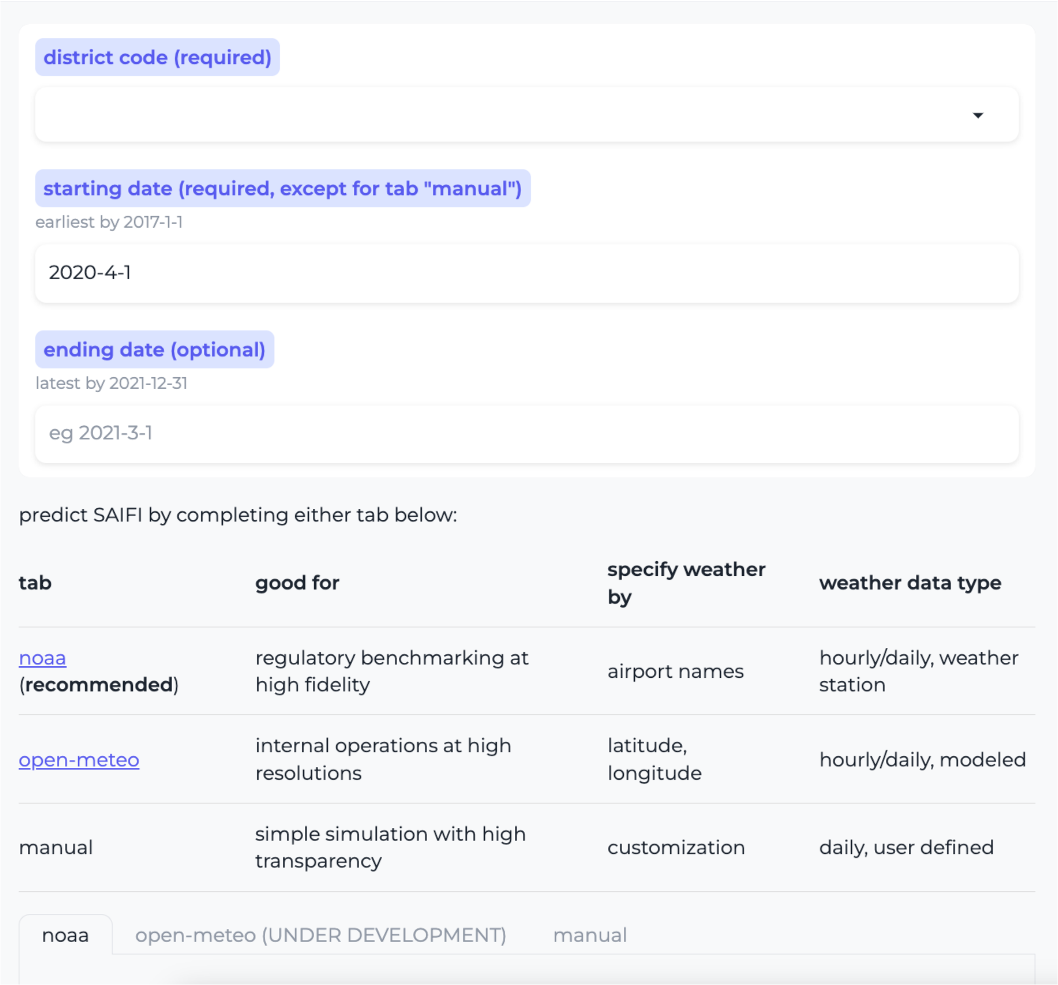
(b)

(c)

(d)

Figure 2: Long-tailed data.
The long-tailed distributions of weather conditions across distribution network district a) a, b) b, and c) c, d) at different outage magnitudes (i.e., System Average Interruption Frequency Index, SAIFI) between 2010-2021.

Figure 3 visualizes an online demo of using our benchmark models. In figure 3.a, we show that two inputs will be required from the user: the district code, and the starting (and ending, if needed) date. Then the user can select one of three usage modes under each of the three tabs, which are tailored for different weather inputs needed for different application cases. Under the "noaa" tab (figure 3.b), the user only needs to specify a list of LCD weather stations, which are mostly airports. These station-based data are highly reliable and accessible, which is ideal for regulatory uses, and our models are developed based on them. For the "open-meteo" option (figure 3.c), which is still under development by the time of publication, we expect the users to enter the latitude and longitude of weather data they are interested in, which are usually sourced from modeled weather grids (e.g., open-meteo.com [39]), and then our server will internally query and process them before feeding them into our pre-trained models. This input mode is particularly useful for the utility's internal operations where spatial (and temporal) resolutions of weather impacts are critical. Last, through the "manual" tab (figure 3.d), the user can manually specify the 8 root weather features, instead of raw weather data as in the previous two modes, that will be directly fed into our models. This design opens up our model with full transparency on feature inputs so that the user can experiment with all kinds of weather scenarios, which is convenient for both emergent storm responses and long-term resilience sandboxing.



district code (required)

starting date (required, except for tab "manual")

earliest by 2017-1-1

2020-4-1

ending date (optional)

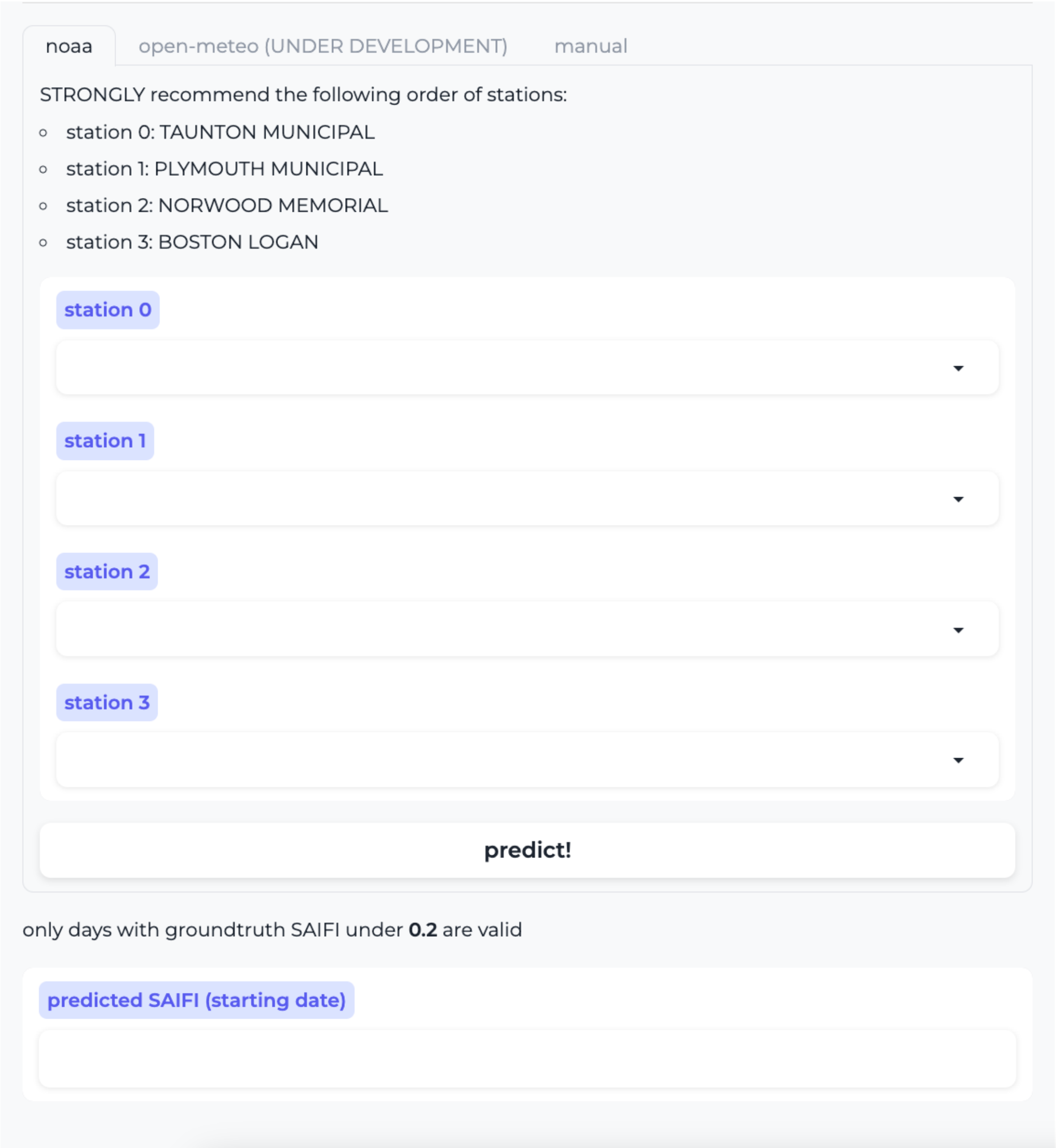
latest by 2021-12-31

eg 2021-3-1

predict SAIFI by completing either tab below:

tab	good for	specify weather by	weather data type
noaa (**recommended**)	regulatory benchmarking at high fidelity	airport names	hourly/daily, weather station
open-meteo	internal operations at high resolutions	latitude, longitude	hourly/daily, modeled
manual	simple simulation with high transparency	customization	daily, user defined

(a)


noaa
open-meteo (UNDER DEVELOPMENT)
manual

STRONGLY recommend the following order of stations:

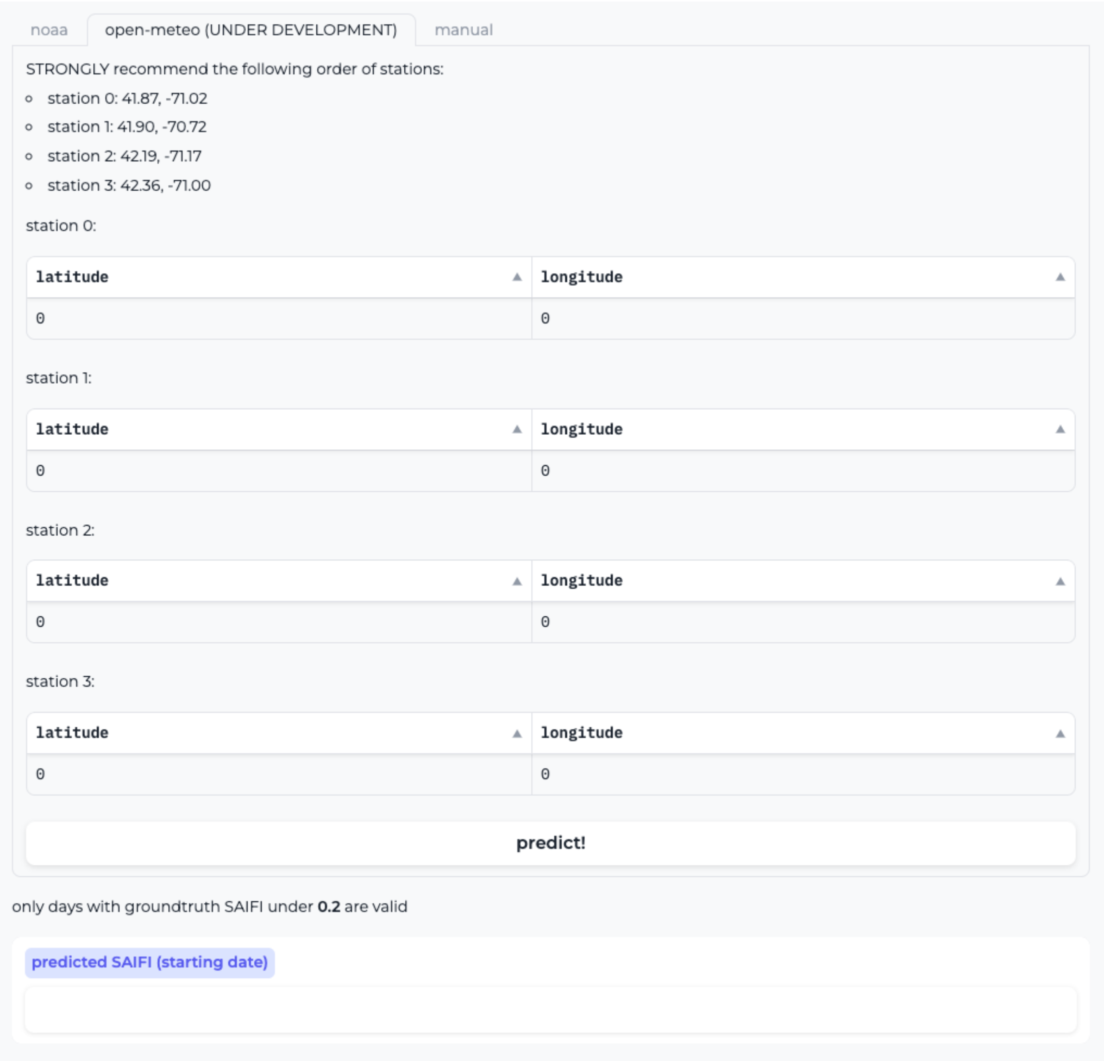
- station 0: TAUNTON MUNICIPAL
- station 1: PLYMOUTH MUNICIPAL
- station 2: NORWOOD MEMORIAL
- station 3: BOSTON LOGAN

station 0
station 1
station 2
station 3
predict!

only days with groundtruth SAIFI under **0.2** are valid

predicted SAIFI (starting date)

(b)



STRONGLY recommend the following order of stations:

- station 0: 41.87, -71.02
- station 1: 41.90, -70.72
- station 2: 42.19, -71.17
- station 3: 42.36, -71.00

station 0:

latitude ▲	longitude ▲
0	0

station 1:

latitude ▲	longitude ▲
0	0

station 2:

latitude ▲	longitude ▲
0	0

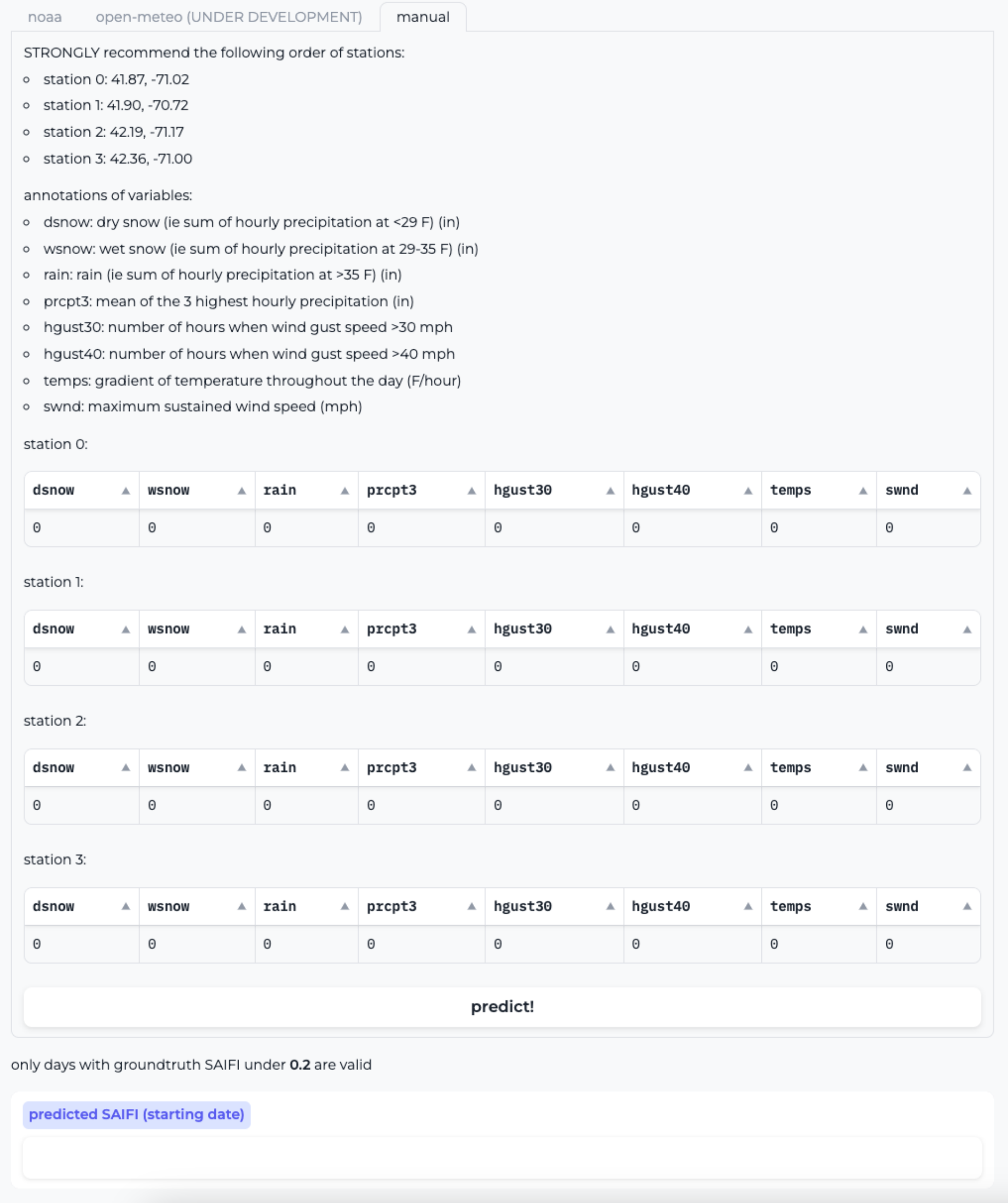
station 3:

latitude ▲	longitude ▲
0	0

only days with groundtruth SAIFI under **0.2** are valid

predicted SAIFI (starting date)

(c)



STRONGLY recommend the following order of stations:

- station 0: 41.87, -71.02
- station 1: 41.90, -70.72
- station 2: 42.19, -71.17
- station 3: 42.36, -71.00

annotations of variables:

- dsnow: dry snow (ie sum of hourly precipitation at <29 F) (in)
- wsnow: wet snow (ie sum of hourly precipitation at 29-35 F) (in)
- rain: rain (ie sum of hourly precipitation at >35 F) (in)
- prcpt3: mean of the 3 highest hourly precipitation (in)
- hgust30: number of hours when wind gust speed >30 mph
- hgust40: number of hours when wind gust speed >40 mph
- temps: gradient of temperature throughout the day (F/hour)
- swnd: maximum sustained wind speed (mph)

station 0:

dsnow	wsnow	rain	prcpt3	hgust30	hgust40	temps	swnd
0	0	0	0	0	0	0	0

station 1:

dsnow	wsnow	rain	prcpt3	hgust30	hgust40	temps	swnd
0	0	0	0	0	0	0	0

station 2:

dsnow	wsnow	rain	prcpt3	hgust30	hgust40	temps	swnd
0	0	0	0	0	0	0	0

station 3:

dsnow	wsnow	rain	prcpt3	hgust30	hgust40	temps	swnd
0	0	0	0	0	0	0	0

only days with groundtruth SAIFI under **0.2** are valid

(d)

Figure 3: Online benchmarking portal.
Design of an online utility resilience benchmarking portal where a) by specifying distribution network districts and dates of interest, users can run our model with inputs tailored for b) regulatory benchmarking, c) utility operations, and d) scenario simulation.

Chapter 5

Conclusion

In this thesis, we review the applications of benchmarking and ML in the regulatory contexts, where we contend that we are methodologically ready to develop ML-based regulatory benchmarks, and potentially integrate them into PIMs to incentivize utility actions. As a case study of applying benchmark-based PIMs, we dive into the definitions and regulations of utility resilience and discuss both the advantages and potential pitfalls of this application, where we argue the benefits exceed the potential drawbacks, especially considering the current regulatory gap on utility resilience under the changing climate. We demonstrate how ML-based regulatory benchmarking could work by developing an RR model using district-level outage and station-level weather data in New England between 2010-2021. Our experiments show the overall promising performances of our ML model as regulatory benchmarks, with future research needed to overcome the long-tailed distributions of outage and weather data. We further illustrate how our model could be scalably deployed and user-friendly accessed for both regulatory and utility usages through an online demo. Lastly, this study will be very relevant within the current policy-making space, where FERC could use our work as a reference case to design a novel resilience standard in response to GAO's recent request [1], and PUCs could explore the explicit regulation of utility resilience by updating their current service-quality PIMs with our weather-adjusted benchmarks.

Bibliography

[1] Climate change is expected to have far-reaching effects and doe and ferc should take actions. US Government Accountability Office, 2021.

[2] John T Abatzoglou. Development of gridded surface meteorological data for ecological applications and modelling. *International Journal of Climatology*, 33(1):121–131, 2013.

[3] International Energy Agency. Electricity security 2021: Climate resilience, 2021.

[4] Per J Agrell and Peter Bogetoft. Theory, techniques, and applications of regulatory benchmarking and productivity analysis. 2018.

[5] Per J Agrell, Peter Bogetoft, et al. Regulatory benchmarking: Models, analyses and applications. *Data Envelopment Analysis Journal*, 3(1-2):49–91, 2017.

[6] Bernard W Bell. Replacing bureaucrats with automated sorcerers? *Daedalus*, 150(3):89–103, 2021.

[7] Luca Bellani, Michele Compare, Enrico Zio, Alessandro Bosisio, Bartolomeo Greco, Gaetano Iannarelli, and Andrea Morotti. A reliability-centered methodology for identifying renovation actions for improving resilience against heat waves in power distribution grids. *International Journal of Electrical Power & Energy Systems*, 137:107813, 2022.

[8] Peter Bogetoft and Anita Eskesen. Balancing information rents and service differentiation in utility regulation. *Energy Economics*, 106:105808, 2022.

[9] Severin Borenstein and James Bushnell. The us electricity industry after 20 years of restructuring. *Annu. Rev. Econ.*, 7(1):437–463, 2015.

[10] Sarah Brody, Matt Rogers, and Giulia Siccardo. Why, and how, utilities should start to manage climate-change risk. *McKinsey & Company. Seattle, WA. https://www. mckinsey. com/industries/electric-power-and-natural-gas/our-insights/why-and-howutilities-should-start-to-manage-climate-change-risk*, 2019.

[11] Joshua W Busby, Kyri Baker, Morgan D Bazilian, Alex Q Gilbert, Emily Grubert, Varun Rai, Joshua D Rhodes, Sarang Shidore, Caitlin A Smith, and

Michael E Webber. Cascading risks: Understanding the 2021 winter blackout in texas. *Energy Research & Social Science*, 77:102106, 2021.

[12] JP Carvallo, FC Hsu, Zeal Shah, and J Taneja. Frozen out in texas: blackouts and inequity. *The Rockefeller Foundation*, 2021.

[13] William C Clark, Ronald B Mitchell, and David W Cash. Evaluating the influence of global environmental assessments. *Global environmental assessments: information and influence. MIT, Cambridge*, pages 1–28, 2006.

[14] Cary Coglianese. The limits of performance-based regulation. *U. Mich. JL Reform*, 50:525, 2016.

[15] Cary Coglianese. Algorithmic regulation. *The Algorithmic Society: Technology, Power, and Knowledge*, 2020.

[16] Kenneth W Costello. Electric power resilience: The challenges for utilities and regulators. *JREG Bulletin*, 37:1, 2019.

[17] William D Eggers, David Schatsky, and Peter Viechnicki. Ai-augmented government. using cognitive technologies to redesign public sector work. *Deloitte Center for Government Insights*, 1:24, 2017.

[18] David Freeman Engstrom and Daniel E Ho. Artificially intelligent government: a review and agenda. *Research Handbook on Big Data Law*, pages 57–86, 2021.

[19] David Freeman Engstrom, Daniel E Ho, Catherine M Sharkey, and Mariano-Florentino Cuéllar. Government by algorithm: Artificial intelligence in federal administrative agencies. *NYU School of Law, Public Law Research Paper*, (20-54), 2020.

[20] Susana Faria and Gilda Soromenho. Fitting mixtures of linear regressions. *Journal of Statistical Computation and Simulation*, 80(2):201–225, 2010.

[21] Chandra Farley, John Howat, Jenifer Bosco, Nidhi Thakar, Jake Wise, and Jean Su. Advancing equity in utility regulation. Technical report, Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States), 2021.

[22] National Centers for Environmental Information. U.s. local climatological data (lcd), 2022.

[23] NARUC Center for Partnerships Innovation. Performance-based regulation state working group (pbrswg) june virtual meeting: Performance incentive mechanisms (pims) for resilience. National Association of Regulatory Utility Commissioners, 2021.

[24] Mireille Hildebrandt. Algorithmic regulation and the rule of law. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2128):20170355, 2018.

[25] Miyuki Hino, Elinor Benami, and Nina Brooks. Machine learning for environmental monitoring. *Nature Sustainability*, 1(10):583–588, 2018.

[26] Earth Networks Inc. Historical earth networks total lightning data (entln), 2021.

[27] T. Jamasb. Benchmarking electricity distribution networks. ACCC/AER Annual Conference, May 2019.

[28] Tooraj Jamasb, Paul Nillesen, and Michael Pollitt. Strategic behaviour under regulatory benchmarking. *Energy Economics*, 26(5):825–843, 2004.

[29] Jennifer Kallay, Alice Napoleon, Jamie Hall, Ben Havumaki, Asa Hopkins, Melissa Whited, Tim Woolf, Jen Stevenson, Robert Broderick, Robert Jeffers, et al. Regulatory mechanisms to enable investments in electric utility resilience. Technical report, Sandia National Lab.(SNL-NM), Albuquerque, NM (United States), 2021.

[30] Chris R. Knittel. Machine learning about outages and weather. MIT CEEPR Webinar on Applying Machine Learning to Energy Policy Research, 2021.

[31] T. Lamn and N. E. Elkind. Clean and resilient: policy solutions for california's grid of the future. Berkeley Law, 2020.

[32] Benjamin D Leibowicz, AH Sanstad, Q Zhu, PH Larsen, and JH Eto. Electric utility valuations of investments to reduce the risks of long-duration, widespread power interruptions, part ii: Case studies. *Sustainable and Resilient Infrastructure*, 8(sup1):203–222, 2023.

[33] Mark Newton Lowry and Lullit Getachew. Statistical benchmarking in utility regulation: Role, standards and methods. *Energy Policy*, 37(4):1323–1330, 2009.

[34] Helen Margetts. Rethinking ai for good governance. *Daedalus*, 151(2):360–371, 2022.

[35] Hila Mehr, H Ash, and D Fellow. Artificial intelligence for citizen services and government. *Ash Cent. Democr. Gov. Innov. Harvard Kennedy Sch., no. August*, pages 1–12, 2017.

[36] Gianluca Misuraca, Colin Van Noordt, et al. Ai watch-artificial intelligence in public services: Overview of the use and impact of ai in public services in the eu. *JRC Research Reports*, (JRC120399), 2020.

[37] Jonas J Monast. Ratemaking as climate adaptation governance. *Frontiers in Climate*, 3:738972, 2021.

[38] Massachusetts Department of Public Utilities. Order adopting revised service quality guidelines. D.P.U. 12-120-D Order, 2015.

[39] open meteo.com. Historical weather api, 2023.

[40] Kenneth Oye. Ids.412 lecture slides. MIT, 2023.

[41] L. Pearl. Sustainability tops utility concerns, as seu survey sees discrepancies on cybersecurity, stranded assets. Utility Dive, 2020.

[42] B Guy Peters. Can we be casual about being causal? *Journal of Comparative Policy Analysis: Research and Practice*, 24(1):73–86, 2022.

[43] Edward Raff. Growing and retaining ai talent for the united states government. *arXiv preprint arXiv:1809.10276*, 2018.

[44] David EM Sappington, Johannes P Pfeifenberger, Philip Hanser, and Gregory N Basheda. The state of performance-based regulation in the us electric utility industry. *The Electricity Journal*, 14(8):71–79, 2001.

[45] David EM Sappington and Dennis L Weisman. Designing performance-based regulation to enhance industry performance and consumer welfare. *The Electricity Journal*, 34(2):106902, 2021.

[46] Nicos Savva, Tolga Tezcan, and Özlem Yıldız. Can yardstick competition reduce waiting times? *Management Science*, 65(7):3196–3215, 2019.

[47] Noelle E. Selin. Ids.410 lecture slides. MIT, 2023.

[48] Gagan Deep Sharma, Anshita Yadav, and Ritika Chopra. Artificial intelligence and effective governance: A review, critique and research agenda. *Sustainable Futures*, 2:100004, 2020.

[49] D. Shea. Performance-based regulation: Harmonizing electric utility priorities and state policy. National Conference of State Legislatures, 2023.

[50] Andrei Shleifer. A theory of yardstick competition. *The RAND journal of Economics*, pages 319–327, 1985.

[51] Renato A Spalding, Luiz HL Rosa, Carlos FM Almeida, Renan F Morais, Marcos R Gouvea, Nelson Kagan, Denis Mollica, Alexandre Dominice, Lucca Zamboni, Gabriel H Batista, et al. Fault location, isolation and service restoration (flisr) functionalities tests in a smart grids laboratory for evaluation of the quality of service. In *2016 17th International Conference on Harmonics and Quality of Power (ICHQP)*, pages 879–884. IEEE, 2016.

[52] Eliza Strickland. Andrew ng: unbiggen ai. IEEE Spectrum, 2022.

[53] Kristo Vaher. Next generation digital government architecture. *Republic of Estonia GCIO Office*, 2020.

[54] Michael Veale and Irina Brass. Administration by algorithm? public management meets public sector machine learning. *Public management meets public sector machine learning*, 2019.

[55] Paul Waidelich, Tomas Haug, and Lorenz Wieshammer. German efficiency gone wrong: Unintended incentives arising from the gas tsos' benchmarking. *Energy Policy*, 160:112595, 2022.

[56] Wenhua Wan, Jianshi Zhao, Eklavyya Popat, Claudia Herbert, and Petra Döll. Analyzing the impact of streamflow drought on hydroelectricity production: A global-scale study. *Water Resources Research*, 57(4):e2020WR028087, 2021.

[57] Romany M Webb, Michael Panfil, and Sarah Ladin. Climate risk in the electricity sector. *Environmental Law*, 51(3):577–666, 2021.

[58] Melissa Whited, Tim Woolf, and Alice Napoleon. Utility performance incentive mechanisms. *A Handbook for Regulators Prepared for the Western Interstate Energy Board*, 2015.

[59] Bernd W Wirtz, Jan C Weyerer, and Carolin Geyer. Artificial intelligence and the public sector—applications and challenges. *International Journal of Public Administration*, 42(7):596–615, 2019.

[60] Vasileios Yfantis, KLIMIS Ntalianis, and FILOTHEOS Ntalianis. Exploring the implementation of artificial intelligence in the public sector: welcome to the clerkless public offices. applications in education. *Journal WSEAS Transactions on Advances in Engineering Education*, 17:76–79, 2020.

[61] Daniel Zhang, Saurabh Mishra, Erik Brynjolfsson, John Etchemendy, Deep Ganguli, Barbara Grosz, Terah Lyons, James Manyika, Juan Carlos Niebles, Michael Sellitto, et al. The ai index 2021 annual report. *arXiv preprint arXiv:2103.06312*, 2021.

[62] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.