Optimizing Heat Alert Issuance with Reinforcement Learning

Ellen M. Considine*[1], Rachel C. Nethery[1], Gregory A. Wellenius[2], Francesca Dominici[1], Mauricio Tec*[1,3]

[1]Department of Biostatistics, Harvard T.H. Chan School of Public Health
[2]Department of Environmental Health, Boston University School of Public Health
[3]Department of Computer Science, Harvard University
*ellen_considine@g.harvard.edu, mauriciogtec@hsph.harvard.edu

Abstract

A key strategy in societal adaptation to climate change is using alert systems to prompt preventative action and reduce the adverse health impacts of extreme heat events. This paper implements and evaluates reinforcement learning (RL) as a tool to optimize the effectiveness of such systems. Our contributions are threefold. First, we introduce a new publicly available RL environment enabling the evaluation of the effectiveness of heat alert policies to reduce heat-related hospitalizations. The rewards model is trained from a comprehensive dataset of historical weather, Medicare health records, and socioeconomic/geographic features. We use scalable Bayesian techniques tailored to the low-signal effects and spatial heterogeneity present in the data. The transition model uses real historical weather patterns enriched by a data augmentation mechanism based on climate region similarity. Second, we use this environment to evaluate standard RL algorithms in the context of heat alert issuance. Our analysis shows that policy constraints are needed to improve RL's initially poor performance. Third, a post-hoc contrastive analysis provides insight into scenarios where our modified heat alert-RL policies yield significant gains/losses over the current National Weather Service alert policy in the United States.

Code —
https://github.com/NSAPH-Projects/heat-alerts_RL
Simulator —
https://github.com/NSAPH-Projects/weather2alert
Extended version (with appendices) —
https://arxiv.org/abs/2312.14196

1 Introduction

Extensive evidence links exposure to extreme heat to increases in morbidity and mortality (Ebi et al. 2021). Heat alerts are a practical and low-cost intervention to mitigate these effects (Ebi et al. 2004) by encouraging protective measures such as hydrating more, avoiding physical exertion outdoors, and opening cooling centers. However, studies investigating the effectiveness of heat alerts have observed mixed results (Weinberger et al. 2018; Wu et al. 2023). Developing methods to optimize the issuance of heat alerts for public health is an open problem.

Various challenges stand in the way of solving this problem. First, issuing too many heat alerts may lead to alert fatigue (Nahum-Shani et al. 2017) and health-protective resource depletion on both individual and community/institutional levels. Sequential decision-making (SDM) methods offer a promising yet unexplored avenue for tackling this issue. Second, despite representing a significant health threat on the population scale, local health impacts of heat—and therefore heat alerts—are small and easily confounded in observational data sets (Weinberger et al. 2021). Rare events and low signal (small effects) have been shown to challenge algorithmic decision-making (Frank, Mannor, and Precup 2008; Romoff et al. 2018). Third, mainstream SDM methods are not suitable for spatially heterogeneous settings in which a single policy is not equally effective in all regions or dynamically changing contexts (Padakandla 2021). However, attempting to identify independent policies for each location drastically reduces the amount of available historical data.

In this paper, we lay the foundation for addressing these challenges by introducing a framework for optimizing heat alert issuance using reinforcement learning (RL). RL allows learning SDM policies for determining when to issue heat alerts, aiming to minimize the population health risk as a negative reward signal. The ultimate vision for this vein of research is deploying data-driven policies to alert the public about extreme heat events (and, eventually, other environmental exposures such as extreme cold or wildfire smoke). In addition to our analysis breaking ground towards this goal, a variety of follow-up work is facilitated by the publication of our SDM environment in an open-source Python package, `weather2alert`[1], compatible with the Gymnasium framework for RL (Towers et al. 2023). This challenging, data-driven environment can be used as a benchmark. Further, our methodology to create this environment can be used as a blueprint for other RL simulators, particularly for problems with exogenous state space components and multiple locations or individuals.

The two main components of our heat alerts RL framework are visualized in Figure 1 and summarized below. First, we create a realistically challenging SDM environment for heat alert issuance, structured as a Markov Decision Process (MDP) under a budget constraint. Key compo-

[1]https://github.com/NSAPH-Projects/weather2alert

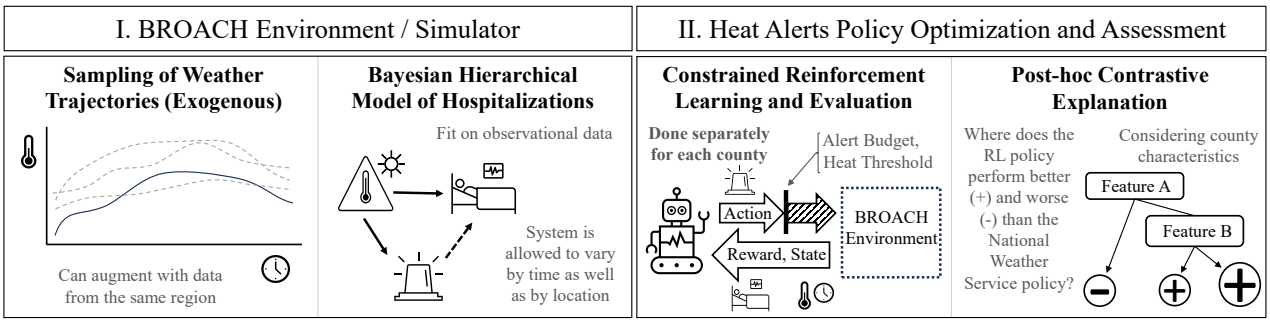

Figure 1: Overview of the heat alerts RL framework.

nents of the dataset used to create this environment are heat alerts issued by the U.S. National Weather Service (NWS), hospitalization records from Medicare, ambient heat index, and other covariates informed by the epidemiological literature. To estimate the rewards under both observed (NWS) and counterfactual alert policies, we fit a statistical model of hospitalizations that allows for spatiotemporal heterogeneity, uncertainty quantification, and fast inference using variational Bayes (Tzikas, Likas, and Galatsanos 2008). For the state-transition model, we exploit a factorization of the state space into an endogenous component with known transition dynamics and an exogenous component sourced from observed weather trajectories. This use of real data allows us to make grounded inferences about the effectiveness of both observed and counterfactual heat alert policies in the U.S. We call this environment structure Bayesian Rewards Over Actual Climate History (BROACH).

Second, we use our environment to train RL models and then to evaluate the RL policies as well as the NWS and alternative baseline policies—all subject to the same alert budget. Our evaluation identifies issues during training that lead to poor performance of standard, widely-used RL algorithms. We address these issues by introducing conceptually simple modifications which enable learning of policies that reduce hospitalizations relative to the NWS policy. These modifications include restricting the RL model to issue alerts only on extremely hot days and optimizing policies for each location separately. Lastly, we perform contrastive explanation (van der Waa et al. 2018; Narayanan, Lage, and Doshi-Velez 2022) of the RL policies by comparing their attributes (e.g., how often does the policy issue alerts on consecutive days) within and across locations to those of the NWS and alternative baseline policies. Specifically, we use visualizations and Classification and Regression Trees (CART) to illustrate systematic differences between these policies and identify characteristics of locations where RL-based policies might offer the greatest improvements in public health.

2 Related Work

Heat alert optimization Recent approaches to improving the issuance of heat alerts include (i) the development of a causal inference technique for stochastic interventions to infer whether increasing the probability of issuing a heat alert would be beneficial (Wu et al. 2023) and (ii) a comparison of methods to identify optimal thresholds above which heat alerts should always be issued (Masselot et al. 2021). Crucially, neither of these approaches addresses the complications of sequential dependence, i.e., the potential for alert fatigue and running out of resources to deploy precautionary measures. To our knowledge, SDM techniques have not yet been explored in a climate and health setting.

RL with exogenous states Several recent papers utilize a decomposition of the state space into exogenous vs. endogenous components, but none of these methods apply directly in our setting. Efroni et al. (2022) provide theoretical analysis and guarantees for sample efficiency in tabular (finite state) environments with an unknown subset of exogenous variables. Sinclair et al. (2023) introduce "hindsight learning" for situations with a known reward function and an exact solver under a fixed exogenous trajectory. Lee et al. (2023) also use hindsight, but do so in the context of identifying latent states within a partially observable MDP. Contrary to Levine, Stone, and Zhang (2024) and Efroni et al. (2024), we do not consider exogenous states as distractors since weather patterns are crucial for decision-making.

Statistical modeling for RL environments Using a statistical model for the reward function in an RL environment is common practice in settings with empirical data. Bayesian models in particular have been used to identify the effect of intervention in health-related applications and dynamic treatment regimes (Liao et al. 2020; Tec, Duan, and Müller 2023; Zajonc 2012). For spatiotemporal or otherwise heterogeneous environments, previous studies have also used a combination of local modeling and global modeling that takes into account the spatial or hierarchical structure to create simulated environments for RL (Wu et al. 2021; Li, Zheng, and Yang 2018; Agarwal et al. 2021).

Constrained learning The use of domain knowledge-based policy restrictions in RL has been found to speed up learning and to guide the learned policy in pragmatic directions (Mu et al. 2021). This supports our restriction of RL heat alerts to very hot days. More generally, incorporating "cost" or "safety" constraints is of interest for many RL applications (Laber et al. 2018; Carrara et al. 2019). A common approach in the constrained RL literature is to use Lagrange multipliers (Guin and Bhatnagar 2023; Ray, Achiam, and Amodei 2019), however, these methods do not enforce

a strict constraint/budget. To enable direct comparison with the observed NWS policy, we impose a strict alert budget, which can be viewed as a simplified variation of the indicator approach formulated by Xu, Zhan, and Zhu (2022). An alternative approach to handling both the alert budget and restriction of alerts to very hot days would be to model the extent of distributional overlap with the observed NWS policy, and penalize actions that have too little overlap (Xu, Zhan, and Zhu 2022). However, our approach has the benefit of being relatively interpretable.

Contrastive policy explanations There is a large literature on post-hoc explainability in RL (Heuillet, Couthouis, and Díaz-Rodríguez 2021). Contrastive analysis has been used to explain differences between RL policies and more familiar / intuitive policies (van der Waa et al. 2018; Narayanan, Lage, and Doshi-Velez 2022). We note that while CART has been used to try and mimic RL policies (Puiutta and Veith 2020), our use of it to analyze differences between policies has not been previously documented.

3 Problem Setup

3.1 RL Preliminaries

RL methods typically operate within the framework of Markov Decision Processes or MDPs (Sutton and Barto 2018). A finite-horizon or episodic MDP with horizon (episode length) H is defined as a tuple $\mathcal{M} = \langle S, A, R, P, d_0, \gamma \rangle$, where S is the set of states, A is the set of actions, $R : S \times A \to \mathbb{R}$ is the expected reward function, $P : S \times A \to \Delta(S)$ is the transition function[2], $d_0 \in \Delta(S)$ is the initial state distribution, and $0 < \gamma \leq 1$ is the time discount factor[3]. A (non-stationary) policy is defined as a collection of decision rules $\pi = \{\pi_t : S \to \Delta(A)\}_{t=0}^{H-1}$, mapping from states to probability distributions over actions at each time step. The state value function is $V_t^\pi(s) := \mathbb{E}_\pi\big[\sum_{h=t}^{H-1} \gamma^{h-t} R(s_h, a_h) | s_t = s\big]$ and the state-action value function is $Q_t^\pi(s,a) := R(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)}[V_t^\pi(s')]$. The policy optimization goal is to identify π^* that maximizes the expected cumulative reward

$$J(\pi) := \mathbb{E}_{s_0 \sim d_0}[V_0^\pi(s_0)]. \quad (1)$$

Under regularity conditions (Puterman 2014), the optimal policy is deterministic and the unique solution (up to ties) of the Bellman optimality equation

$$Q_t^*(s,a) = R(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)}[\max_{a'} Q_{t+1}^*(s', a')], \quad (2)$$

where Q_t^* denotes the state-action value function of π_t^*. The optimal action satisfies $\pi_t^*(s) = \arg\max_a Q_t^*(s,a)$.

RL algorithms search for the optimal policy of an MDP by interacting with an environment that generates rewards and transitions, where the RL algorithm does not need knowledge of the transition function (Sutton and Barto 2018).

[2] $\Delta(\cdot)$ is the set of probability distributions over the set "$\cdot$".

[3] Note that in our analysis, $\gamma = 1$ works best, so in practice there is no discounting.

3.2 Issuing Heat Alerts as a Constrained MDP

The heat alerts MDP components are summarized here. See also Table S1 of Appendix A. Each MDP $\mathcal{M}_k$ corresponds to a geographic area $k \in \mathcal{K}$, and episodes are indexed as $j \in \mathcal{J}$. In this work, geographic areas are U.S. counties, and each episode spans the warm season / summer from May 1st to September 30th of a specific year ($H = 152$ days).

The action a_t on day t is either 1 (issue a heat alert) or 0 (do not issue a heat alert). To mitigate alert fatigue, we adopt a strict action budget-constrained approach, optimizing the policy subject to $\sum_{t=0}^{H-1} a_t \leq b$.

The reward $r_t = R(s_t, a_t)$ is the expected rate of heat-related hospitalizations at time t, transformed such that fewer hospitalizations correspond to a greater reward. Specifically, if $\rho_t(a_t) := \rho(s_t, a_t)$ represents the *per capita* rate of heat-related hospitalizations on day t if action $a_t \in \{0, 1\}$ is taken at state s_t, then $r_t \propto -\rho_t(a_t) + \text{const}$.

The state vector s_t contains the factors underlying the effect of heat on hospitalizations and the effectiveness of heat alerts at reducing hospitalizations. It can be decomposed as $s_t = (\xi_t, x_t)$, where ξ_t is the exogenous component and x_t is the endogenous component. The exogenous component is defined as the portion of the state space that is not influenced by the agent's actions a_t, such as the weather. The transition function of the endogenous component (heat alert history) is known and deterministic. Hence, the full state transition function admits a factorization

$$\begin{aligned} P(s_{t+1}|s_t, a_t) &= P((\xi_{t+1}, x_{t+1})|(\xi_t, x_t), a_t) \\ &= P_\xi(\xi_{t+1}|\xi_t) P_x(x_{t+1}|x_t, a_t) \end{aligned} \quad (3)$$

4 BROACH: An RL Environment for Optimizing Heat Alert Issuance

In creating an interactive heat alert environment, our design objectives are (i) allow use of general-purpose RL algorithms, (ii) be grounded in real data, and (iii) handle the low signal and heterogeneity of health impacts of heat and heat alerts across space and time. To meet these objectives, we introduce a methodology termed Bayesian Rewards Over Actual Climate History (BROACH). BROACH can be generalized to other climate-related events and interventions.

4.1 Data Sources

Here we provide an overview of the data used in the RL environment; Appendix B contains more details.

Heat alerts and heat index We use the dataset of Weinberger et al. (2021) with daily, county-level records (2006-2016, May-September) of heat alerts issued by the NWS, as well as heat index, which is a measure of the combined effect of temperature and humidity. To understand the observed heat alert data, note that while the decision to issue an alert is based on temperature thresholds, it is also strongly affected by the discretion of the local office (Hawkins, Brown, and Ferrell 2017). Analysis by (Hondula et al. 2022) suggests that spatial variability in the current NWS/local office approach is not well aligned with the health risk from heat.

Heat-related hospitalizations We merged the heat alert and weather data with daily, county-level hospitalization counts from Medicare. We included cause-specific hospitalizations previously found to be associated with extreme heat in the Medicare population: septicemia, peripheral vascular disease, urinary tract infections, and diabetes mellitus with complications (Bobb et al. 2014). We excluded heat stroke and fluid and electrolyte disorders (which were also found to be associated with extreme heat in Medicare) because Weinberger et al. (2021) observed a positive association between heat alerts and hospitalizations with these causes. They hypothesized it was due to increased awareness of heat-related symptoms and seeking medical care. Figure S1 illustrates this scenario as an unobserved mediation problem (Pearl 2012), which affects the identification of the preventive effect of heat alerts. We refer to the remaining causes, which we pool together into one outcome variable, as not-obviously heat-related (NOHR) hospitalizations.

Counties We consider the 761 counties with a population greater than 65,000 to avoid locations with very few hospitalizations and to focus on the most populous areas. Figure 2 shows the counties considered, which account for approximately 75% of the population and 25% of the number of counties.

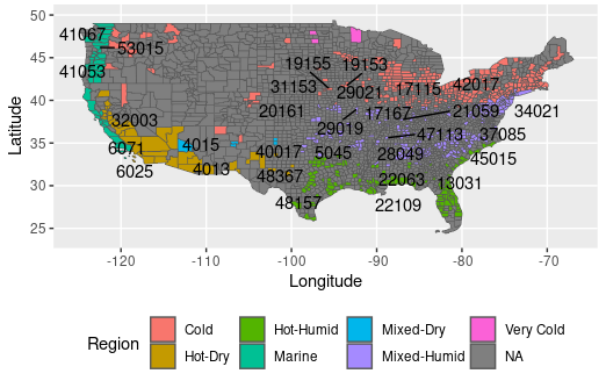

Figure 2: Map of the counties considered and their regional climate zone classifications. All the colored counties are used in the Bayesian rewards model and RL environment; the 30 counties with annotated FIPS codes are used in the RL experiments.

County-level characteristics Spatial heterogeneity in the health effects of heat and heat alerts can occur due to people in different regions being affected differently by the same absolute temperatures due to geographical self-selection and climate adaptation, differential health susceptability and/or individual agency associated with socioeconomic status and population density, and variable response to heat alerts due to local policy and political ideology (Ng et al. 2014; Zanobetti et al. 2013; Errett et al. 2023; Cutler et al. 2018). To characterize this spatial heterogeneity, we compiled a set of county-level covariates: population density and median household income (U.S. Census Bureau 2014), regional classifications of climate zones (U.S. Energy Information Administration 2020), broadband (internet) usage data (Microsoft AI for Good Research Lab 2021), presidential election returns (MIT Election Data and Science Lab 2018), and fine particulate matter ($PM_{2.5}$) (Hammer et al. 2020). For the latter, there is strong evidence of adverse synergistic health impacts of air pollution and heat (Anenberg et al. 2020).

4.2 The State Space

Recall that the state space can be factorized as $s_t = (\xi_t, x_t)$, where ξ_t is the exogenous component and x_t is the endogenous component. The exogenous component contains the quantile value of the day's observed heat index (QHI)—within each county across warm seasons for all the years—which allows us to account for long-term climate adaptation. It also contains time-varying factors which can modify the heat-health relationship (Heo, Bell, and Lee 2019; Anderson and Bell 2011): day of summer (which is equivalent to t), weekend status, and excess heat compared to the last three days. The endogenous component contains the number of heat alerts issued in the last 14 days, an indicator of whether a heat alert was issued yesterday, and the remaining alert budget for that episode.

4.3 The Reward Function: A Hierarchical Model of Hospitalizations

To learn the expected hospitalization rate, and thereby the transformed reward r_t, we train a Bayesian hierarchical model with the desiderata of (a) facilitating obtainment of confidence intervals when evaluating RL policies and (b) allowing for spatial heterogeneity in the health effects of heat and heat alerts. Let $\rho_t^{(k,j)}(a)$ be the expected per capita NOHR hospitalization rate at time t in county-summer (k, j) when taking action a. We propose the scaled reward function $r_t^{(k,j)}(a) := C_1 - C_2\rho_t^{(k,j)}(a)$ where C_1 and C_2 are positive scaling constants. (Note that choices of C_1 and C_2 do not change the optimal policy, but can help stabilize learning.) We choose $C_1 = 1$ and C_2 as the reciprocal of the observed hospitalization rate through the entire summer.

Let $y_t^{(k,j)}(a)$ denote the counterfactual outcome representing the number of NOHR hospitalizations for action $a \in \{0, 1\}$, and $n^{(k,j)}$ denote the total population susceptible to hospitalization in county k in episode j. We introduce two functions—to be learned—governing the counterfactual hospitalization rate in county k as a function of $s_t^{(k,j)}$. First, $\lambda_k \colon S \to \mathbb{R}^+$ denotes the baseline rate of hospitalizations when no alert is issued. Second, $\tau_k \colon S \to (0, 1)$ expresses the effectiveness of issuing an alert as a multiplicative reduction factor. The top-level hospitalization model is

$$\begin{aligned} y_t^{(k,j)}(a) &\sim \text{Poisson}(n^{(k,j)}\rho_t^{(k,j)}(a)), \\ \rho_t^{(k,j)}(a) &:= \lambda_k(s_t^{(k,j)})(1 - a \cdot \tau_k(s_t^{(k,j)})), \end{aligned} \tag{4}$$

for all $(k, j) \in \mathcal{K} \times \mathcal{J}$ and $t \in \{0, \ldots, H - 1\}$. The Poisson distribution/loss function is the natural choice for count data and recommended for health outcomes data due to its correspondence to the Cox survival model (Austin 2017).

To estimate the functions λ_k and τ_k under limited data and low health signal, we incorporate a data-driven random effects prior (using the county-level features) and inject domain knowledge on the signs of certain coefficients. We train this model using variational inference, using the Python package Pyro (Harpole et al. 2019). **The details of this approach, as well as model diagnostics, are in Appendix C.** Note that the fitted coefficients in the alert effectiveness component indicate the presence of alert fatigue. To utilize this rewards model within the BROACH environment, we draw from the variational posterior distribution at the start of each episode, and then use those coefficients to calculate $r_t^{(k,j)}$ given $\left(s_t^{(k,j)}, a_t^{(k,j)}\right)$ for $t = 0, ..., H-1$.

4.4 The Alert Budget

Each episode in our MDP starts with a fixed alert budget b. This can be specified as a constant, sampled from a range of values, or set to the observed (NWS) value. As the agent steps through the episode, once $\sum_{h=0}^{t} a_h = b$, it is prohibited from issuing any more alerts: it can only obtain the reward $r_h^{(k,j)}(0)$ for days $h = t+1, ..., H-1$.

4.5 The Transition Function: Observing Actual Weather/Climate History

We leverage the decomposition equation 3 to use real data for the exogenous trajectories, and introduce a data augmentation scheme in the context of single-county RL.

Sampling of exogenous trajectories Recall that the endogenous aspect of our state space, heat alert history, is deterministically updating, so does not need to be modeled probabilistically. The key observation in the heat alert setting is that the more complicated aspect of the environment, the weather/climate, also does not need to be modeled because it is completely exogenous to heat alert decision-making. In this setting, developing a model for the weather is not only unnecessary, but would unavoidably introduce error. Instead, we sample real observed weather trajectories.

Regional data augmentation The issue with this approach is that there are only 11 years of exogenous trajectories (2006-2016) available per county during which we also have heat alert data—which is needed when we use observed alert budgets (paired with observed weather) for comparison with the NWS. To mitigate the potential for RL overfitting to these 11 years, we propose data augmentation by sampling exogenous trajectories from other counties in the same regional climate zone[4] (U.S. Energy Information Administration 2020) during RL training and validation / tuning.

5 Learning and Evaluation

We conduct experiments to (a) assess the ability of standard RL algorithms to learn effective heat alert policies in the BROACH environment, (b) test various modifications to these algorithms, and (c) compare the resulting policies to the NWS policy and other intuitive baselines. We then investigate heterogeneous performance of the RL policies to provide domain-relevant insights. Code for all steps is available in our GitHub repository[5].

5.1 Policy Constraint: Very Hot Days

Foreshadowing our findings, standard RL algorithms struggle to learn to conserve their alert budgets for later in the summer, resulting in poor performance. A major modification we implement to encourage more effective behavior is restricting the issuance of heat alerts to days above a QHI threshold, optimized separately for each county. To optimize the QHI threshold, we test the sequence of values between 0.5 and 0.9 (ensuring overlap with the NWS policy), by every 0.05, and select the value that yields the best return on our validation set (specified in the following section). Names of models including this adaptation have the suffix ".QHI".

5.2 Experimental Setup

For all training and evaluations, to directly compare counterfactual alert policies with the NWS policy, we fix b in each county-summer to the observed number of heat alerts.

RL baselines We investigate using four common RL algorithms: Deep Q-learning (DQN), Quantile Regression Deep Q-learning (QRDQN), Trust Region Policy Optimization (TRPO), and Advantage Actor-Critic (A2C)—additional information on these algorithms is in Appendix D.2. We use standard implementations of these methods available in the `Stable-Baselines3` Python library (Raffin et al. 2021). To ground our analysis and discussion, note that DQN and QRDQN are off-policy methods which learn deterministic policies by directly solving for the optimal value function using the Bellman optimality condition in equation 2; exploration is induced only during training (e.g. epsilon-greedy). Whereas, TRPO and A2C are on-policy methods which learn stochastic policies by direct optimization of the expected return in equation 1, using refinements of the policy gradient theorem (Sutton et al. 1999); exploration is inherent.

NWS and simple alternative baselines To evaluate each of the RL policies, we compare them to the observed NWS policy as well as several counterfactual baselines: most simplistically, randomly selecting b days on which to issue alerts (RANDOM) and selecting the b days with the highest QHI that summer (TOPK)—the latter is an oracle policy (not implementable in the real world) because to use it, we would have to know the whole summer's daily QHI in advance. We also implement the general guidelines that NWS recommends in the absence of local criteria for issuing heat alerts (BASIC.NWS): alert if the heat index is $\geq 100°$F in northern states and $\geq 105°$F in southern states (Hawkins, Brown, and Ferrell 2017). Lastly, we implement a policy of always issuing alerts on days above a per-county optimized QHI threshold until the budget runs out (AA.QHI).

[4] Due to the large size/geographic range of the "Cold" zone, we grouped its counties into eastern and western subsets.

[5] https://github.com/NSAPH-Projects/heat-alerts_RL

Training data The counties for which NWS issued few heat alerts had high-variance estimates of heat alert effectiveness in the rewards model. Therefore, for the RL experiments, we selected a subset of 30 counties in which at least 75 heat alerts were issued during 2006-2016; spread across the five major climate regions, prioritizing those with higher variance in estimated alert effectiveness across days. See map in Figure 2 and Appendix D.1 for more details.

Evaluation metric The main metric we use for policy comparison is the average cumulative reward per episode ("average return"). To estimate this metric fairly, we hold three years of data (2007, 2011, and 2015) out of RL training, referred to as the evaluation years[6]. For the purposes of RL hyperparameter tuning (details in Appendix D.3) and selection of the optimal QHI threshold for each county, we consider the regionally augmented weather trajectories and associated heat alert budgets from the evaluation years *excluding the county of interest* to be the validation data set. The final evaluation results (reported in tables and figures) are obtained by drawing 1,000 sets of coefficients from the Bayesian rewards model posterior and calculating the return under each policy with weather and budgets *only from the county of interest* during the evaluation years. These validation/tuning and evaluation procedures are described algorithmically in Appendix E. After calculating the average returns for the 30 counties under each competing policy, we compare each policy's returns with the associated returns under the NWS policy using a Wilcoxon-Mann-Whitney test. This nonparametric test allows us to compare the distribution of differences in returns, accounting for the fact that the counties are highly heterogeneous, so the differences between the policies are not normally distributed.

Sensitivity analysis In the main analysis, we evaluate each county-specific implementation of DQN, QRDQN, TRPO, and A2C by allowing each algorithm to deploy the type of policy that it optimizes during training: deterministic for DQN and QRDQN and stochastic for TRPO and A2C. As a sensitivity analysis, we also deterministically evaluate the TRPO and A2C policies, selecting whichever action has a higher probability under the policy function. The names of these models have the prefix "DET". In another sensitivity analysis, we investigate whether the RL models can learn to anticipate subseasonal variation in the warm season by augmenting the RL state space with information about the future. Specifically, we test whether inclusion of the change in heat index over each of the next 10 days, as well as the 50th-100th (by every 10) percentiles of QHI for the remainder of the summer, improves RL performance. In real-time deployments of RL models, this kind of future information would not be known but could be sourced from weather forecasts or climate model projections. Of course, such forecasts / model projections are not perfect, but we start with perfect future information as a proof of concept. The names of models that include future information have the suffix ".F".

[6]Appendix D.4 shows the results of a sensitivity analysis where the models are trained on 2006-2013 and evaluated on 2014-2016.

5.3 Results and Discussion

Figures 3, S5, and S6 illustrate how the different policies look in practice, respectively in scenarios where using RL is beneficial, where the alert budget is small, and where using RL is not beneficial compared to the alternatives.

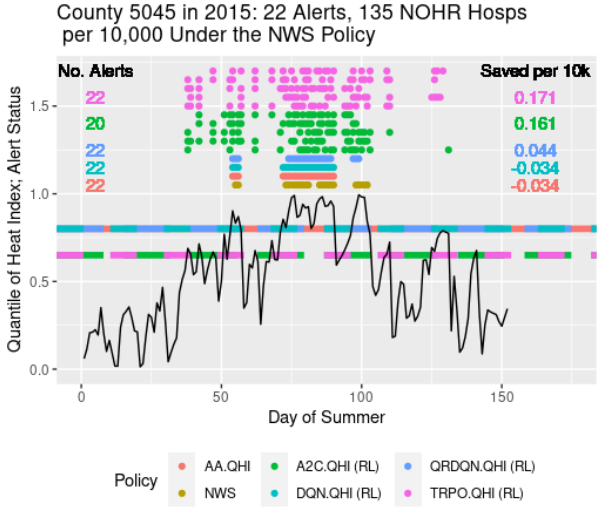

Figure 3: An example of observed and counterfactual heat alert policies for a single summer (2015 is the most recent year in our evaluation set), with estimates of the number of NOHR hospitalizations saved (compared to NWS) per 10,000 Medicare enrollees under each policy. The dashed lines indicate the optimized QHI threshold of the policy in the same color. The multiple horizontal lines of pink and green dots indicate five different samples from TRPO.QHI and A2C.QHI respectively. For these two policies, the number of NOHR hospitalizations saved is the average of all their evaluations over 2015.

Table 1 provides the main results of our RL experiments. (An extended version, including absolute number of hospitalizations, can be found in Table S5.)

QHI restriction is necessary for the standard RL algorithms to perform well Several counterfactual policies perform worse than NWS (as indicated by negative median differences in returns): RANDOM, BASIC.NWS, and the RL models without the QHI restriction. By contrast, the RL models with the QHI restriction perform significantly better than NWS, as indicated by their significant positive median difference in returns. A rough estimate if A2C.QHI were implemented as-is across all counties in the U.S. (using the Medicare population from 2011, the midpoint of our study period) is that we could see a reduction of 222 NOHR hospitalizations per year (approximate 95% CI = (-491, 1131); details of this CI calculation are in Appendix F). Discussion of this absolute benefit to public health is in Appendix G.

The futility of including future information Across both plain and QHI-restricted models, TRPO was the only one that benefited from including future information in the state

Policy	Median Difference	WMW	P-value
*TOPK	0.022	406	0.0002
RANDOM	-0.015	177	0.88
BASIC.NWS	-0.286	30	1.0
AA.QHI	0.03	348	0.0025
DQN	-0.123	43	0.99995
QRDQN	-0.117	51	0.99991
TRPO	-0.065	97	0.99742
A2C	-0.063	100	0.99689
DQN.QHI	0.03	370	0.00242
QRDQN.QHI	0.035	344	0.01121
TRPO.QHI	0.038	338	0.0154
A2C.QHI	**0.042**	344	0.01121
DQN.F	-0.417	40	0.99996
QRDQN.F	-0.42	48	0.99993
TRPO.F	-0.062	103	0.99625
A2C.F	-0.063	101	0.99669
DQN.QHI.F	-1.895	1	1
QRDQN.QHI.F	0.014	238	0.45904
TRPO.QHI.F	**0.046**	345	0.01062
A2C.QHI.F	0.036	343	0.01183
DET.TRPO.QHI	-0.662	57	0.99985
DET.A2C.QHI	-0.886	90	0.99837
DET.TRPO.QHI.F	**0.032**	271	0.21723
DET.A2C.QHI.F	0.02	287	0.13335

Table 1: Comparison between the average return of each counterfactual policy and that of the NWS policy on the evaluation years, summarized across counties (e.g. "Median Diff." is the median difference in average return). WMW is the Wilcoxon-Mann-Whitney statistic (higher is better); its associated p-value is also included. The first policy, marked by *, requires oracle knowledge of the future weather.

space. While A2C was minimally impacted, QRDQN and DQN were made noticeably worse. The latter may be due to the future information dramatically increasing the size of the state space/neural network parameters, which our hyperparameter tuning did not fully counteract. Note that for the TRPO models, the benefit observed with the use of future information would likely be smaller in practice due to forecast / prediction error—the simulation of which is beyond the scope of this paper. Therefore, we focus on the RL models without future information in the remainder of this text.

The need for stochastic policies Interpreting policies from the TRPO and A2C models deterministically fails in the heat alerts setting under our rewards model. Additional discussion of this result is in Appendix H.2. This finding raises the question: *Are stochastic policies palatable from a domain perspective?* On the one hand, it is less immediately satisfying that some aspect of heat alerts issuance would be left to chance. On the other hand, in the context of human-in-the-loop decision making—which would likely be the case for public organizations such as NWS (Stuart et al. 2022)—an algorithm reporting probabilities is more informative than reporting only a binary action. In any case, if in the future a heat alerts RL model was running continuously online, it would likely need to utilize exploration (incorporating some randomness into its actions) to update itself over time.

Comparing policies that perform significantly better than NWS We note that the performance and behavior of DQN.QHI is practically identical to AA.QHI—not just in its numerical summary, but also in its policy behavior. QRDQN.QHI performs a bit better than DQN.QHI, but not as well as the stochastic policies. Between those, A2C.QHI performs better than TRPO.QHI, though their policies tend to display similar characteristics. Therefore, we focus on A2C.QHI as the best RL policy in the post-hoc analysis. The non-RL counterfactual policies that perform well are that of always issuing an alert on days above an optimized QHI threshold (AA.QHI) and of issuing alerts on the b hottest days of the summer (TOPK). We note that the ordering of the Wilcoxon-Mann-Whitney statistics for TOPK, AA.QHI, and A2C.QHI is opposite the ordering of these policies' median difference in returns (compared to NWS). This is due to the longer tails of the A2C.QHI differences, as illustrated in Figure S8. For interpreting these results, recall that TOPK is an oracle policy. However, AA.QHI is an implementable alternative to A2C.QHI, so we must consider it more seriously.

5.4 Post-hoc Contrastive Explanation

To characterize differences between the best RL policies, the NWS policy, and alternative baselines, we consider both stationary features included upstream in the analysis (e.g., climate region and median household income) as well as characteristics of both the NWS and counterfactual alert policies, namely the distributions of the days of summer on which alerts are issued and the lengths of alert streaks (sequences of repeated alerts). We start with these alert characteristics, descriptive histograms of which are in Figure S7. Both A2C.QHI and AA.QHI tend to issue heat alerts earlier in the summer than NWS. A2C.QHI also tends to issue shorter streaks of alerts (i.e., fewer consecutive days of alerts), which makes sense given its ability to learn sequential dependence encoded by the rewards model.

Figure S8 illustrates the varying performance of TOPK, AA.QHI, and A2C.QHI relative to the NWS policy, across climate regions and counties. Note that large heterogeneity in the health effects of heat alert intervention across counties is consistent with previous findings (Wu et al. 2023). A striking visual pattern is the increasing vertical spread from left to right: TOPK performs the least heterogeneously across counties, followed by AA.QHI, and finally A2C.QHI. Similarly, the humid climate regions display more heterogeneity across counties than the others, across all the policies. This is partially due to those regions being larger and thus overrepresented among our 30 counties (see Table S3), but that does not fully explain the discrepancy.

To structure our investigation of why some counties see smaller (or even no) benefits from the application of the heat alerts RL, we fit CART on both the numeric difference in average return (compared to NWS) and a polytomous indicator of which policy performed the best. To prevent CART overfitting to our 30 counties and to facilitate interpretation, we ensure that the minimum number of counties in each leaf (terminal) node is greater than or equal to five for the regression and three for the polytomous classification.

RL performs best in counties with larger heat alert-health signal and longer heat waves A classification tree distinguishing A2C.QHI, AA.QHI, NWS, and TRPO.QHI is shown in Figure S9. The most predictive feature of the counterfactual policies' performance relative to NWS is the size of the alert budget. The fact that the other policies find it harder to improve on NWS with more alerts may be because in such settings it is less critical to discriminate days when heat alerts will be most effective, given that our rewards model assumes more alerts never hurt in an absolute sense. The second split in the tree indicates that the RL performs best when there is greater variation in alert effectiveness across days, in other words, when there is more signal in the heat alert-health relationship that can be leveraged by the RL. The third split highlights how it is better to use RL in counties that tend to experience more prolonged heat waves, which cause AA.QHI to issue more alerts in a row, increasing the likelihood of alert fatigue. Running the same polytomous classification on a set of more conventional covariates (Figure S10), we see a nearly-identical tree except with median household income in place of alert effectiveness and humidity in place of prolonged heat waves. Both of these associations make sense from a domain science perspective.

Additionally, RL performs better than NWS earlier in the summer Conducting regression on the difference in average returns between A2C.QHI and NWS using CART (shown in Figure S11) generated complementary insights. The most predictive feature is the median day of summer on which the RL issued heat alerts: A2C.QHI performs better in counties for which it identified that it is optimal to issue alerts earlier in the summer. Whereas, in counties for which it is optimal to issue alerts later in the summer, there was less room for improvement by the RL because the NWS already tends to issue alerts later in the summer.

6 Conclusion and Future Work Directions

This work lays the foundation of SDM for climate & health, by (1) formulating and building an SDM pipeline to optimize heat alert issuance, via integrating traditional statistical methods with cutting-edge RL techniques, and (2) offering insights into scenarios where RL improves vs. fails to improve on the observed policy or simpler alternatives.

Ultimately, we found that it was necessary to restrict alerts to days with higher QHI for standard RL algorithms (DQN, QRDQN, TRPO and A2C) to be effective. Our post-hoc analysis enabled investigation of where RL performed better than the NWS policy and the simpler alternative AA.QHI. Future work might explore the use of methods which ensure a new policy is never worse than the existing policy, such as safe policy learning or model predictive control. Alternatively, we could develop a preliminary predictive model to select counties that are likely to benefit from RL, using intuitions similar to those in our CART analysis.

We anticipate several common questions about our approach. The first is how to move beyond a fixed alert budget, important in a changing system (under climate change, the number of extreme heat events is expected to increase). From a practical standpoint, real monetary budgets from stakeholders could be used, for instance "how many times can we afford to open our cooling center(s) this summer?" Future work could also explore modeling alert fatigue as part of the RL environment, for instance by incorporating tools from behavioral science. RL algorithms that are intentionally robust to distribution shift also merit investigation in a climate & health setting.

A second question is whether offline (batch) RL methods, which use only observed data in place of a simulator, could be used to circumvent dependence on the specification of a rewards model. Several major challenges stand in the way of an offline approach, highlighted in our heat alerts example. First, there is a tension between trying to improve the observed policy and controlling distribution shift (Levine et al. 2020): for instance, we see that NWS often issues alerts in streaks, making it hard for offline RL to learn and optimize the impact of an individual alert. Second, assessing the performance of offline RL using off-policy evaluation methods is made difficult by long episodes, large potential for distribution shift (especially in the absence of a modification such as restricting alerts to very hot days), and high degree of autocorrelation in the reward (Uehara, Shi, and Kallus 2022).

We close with a broader view of climate & health decision making. Ideally, an optimal heat alert system would account for different types of health impacts experienced by different demographic groups—future work could explore multi-objective RL. Similarly, it is important to recognize that the issuance of heat alerts is only a first step in reducing the public health impacts of extreme heat (Errett et al. 2023): there is much more work to be done analyzing and expanding local actions in response to heat alerts. If such on-the-ground interventions are able to increase the effectiveness of heat alerts, then this is likely to increase the ability of SDM methods such as RL to identify these effects and help us continue improving our strategies in the future.

Acknowledgements

This work was supported by an NSF Graduate Research Fellowship (EMC), NIH award K01ES032458 (RCN), NIH award R01-ES029950 (MT, GAW), NIH award 1R01MD016054-01A1 (FD), NIH award 5R01ES030616-04 (MT), NIH award 3RF1AG074372-01A1S1 (MT), NIH award P30ES000002 (MT), Alfred P. Sloan Foundation grant G-2020-13946 (FD), the Fernholz Foundation (MT), and Wellcome Trust grant 216033-Z-19-Z (GAW). We thank the National Studies on Air Pollution and Health research group for their support of this project. The computation in this paper was performed on the FASSE and FASRC Cannon clusters supported by the FAS Division of Science Research Computing at Harvard University. Lastly, we thank Kate Weinberger for facilitating access to the heat alerts data set, and the STAT 234 teaching staff at Harvard University for their feedback during early stages of this project.

Disclosures: GAW currently serves as a consultant for the Health Effects Institute (Boston, MA) and recently served as a consultant for Google, LLC (Mountain View, CA).

References

Agarwal, A.; Alomar, A.; Alumootil, V.; Shah, D.; Shen, D.; Xu, Z.; and Yang, C. 2021. PerSim: Data-Efficient Offline Reinforcement Learning with Heterogeneous Agents via Personalized Simulators. In *Advances in Neural Information Processing Systems*, volume 34, 18564–18576. Curran Associates, Inc.

Anderson, G. B.; and Bell, M. L. 2011. Heat Waves in the United States: Mortality Risk during Heat Waves and Effect Modification by Heat Wave Characteristics in 43 U.S. Communities. *Environmental Health Perspectives*, 119(2): 210–218.

Anenberg, S. C.; Haines, S.; Wang, E.; Nassikas, N.; and Kinney, P. L. 2020. Synergistic health effects of air pollution, temperature, and pollen exposure: a systematic review of epidemiological evidence. *Environmental Health*, 19: 130.

Austin, P. C. 2017. A tutorial on multilevel survival analysis: methods, models and applications. *International Statistical Review*, 85(2): 185–203.

Bobb, J. F.; Obermeyer, Z.; Wang, Y.; and Dominici, F. 2014. Cause-Specific Risk of Hospital Admission Related to Extreme Heat in Older Adults. *JAMA*, 312(24): 2659–2667.

Carrara, N.; Leurent, E.; Laroche, R.; Urvoy, T.; Maillard, O.-A.; and Pietquin, O. 2019. Budgeted reinforcement learning in continuous state space. *Advances in Neural Information Processing Systems*, 32.

Cutler, M. J.; Marlon, J. R.; Howe, P. D.; and Leiserowitz, A. 2018. The Influence of Political Ideology and Socioeconomic Vulnerability on Perceived Health Risks of Heat Waves in the Context of Climate Change. *Weather, Climate, and Society*, 10(4): 731–746.

Ebi, K. L.; Capon, A.; Berry, P.; Broderick, C.; de Dear, R.; Havenith, G.; Honda, Y.; Kovats, R. S.; Ma, W.; Malik, A.; Morris, N. B.; Nybo, L.; Seneviratne, S. I.; Vanos, J.; and Jay, O. 2021. Hot weather and heat extremes: health risks. *The Lancet*, 398(10301): 698–708.

Ebi, K. L.; Teisberg, T. J.; Kalkstein, L. S.; Robinson, L.; and Weiher, R. F. 2004. Heat Watch/Warning Systems Save Lives: Estimated Costs and Benefits for Philadelphia 1995–98. *Bulletin of the American Meteorological Society*, 85(8): 1067–1074.

Efroni, Y.; Foster, D. J.; Misra, D.; Krishnamurthy, A.; and Langford, J. 2022. Sample-efficient reinforcement learning in the presence of exogenous information. In *Conference on Learning Theory*, 5062–5127.

Efroni, Y.; Misra, D.; Krishnamurthy, A.; Agarwal, A.; and Langford, J. 2024. Provable RL with exogenous distractors via multistep inverse dynamics. *ICML Workshop on Reinforcement Learning Theory*.

Errett, N. A.; Hartwell, C.; Randazza, J. M.; Nori-Sarma, A.; Weinberger, K. R.; Spangler, K. R.; Sun, Y.; Adams, Q. H.; Wellenius, G. A.; and Hess, J. J. 2023. Survey of extreme heat public health preparedness plans and response activities in the most populous jurisdictions in the United States. *BMC public health*, 23(1): 811.

Frank, J.; Mannor, S.; and Precup, D. 2008. Reinforcement learning in the presence of rare events. In *Proceedings of the 25th international conference on Machine learning*, 336–343.

Guin, S.; and Bhatnagar, S. 2023. A policy gradient approach for finite horizon constrained Markov decision processes. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, 3353–3359. IEEE.

Hammer, M. S.; van Donkelaar, A.; Li, C.; Lyapustin, A.; Sayer, A. M.; Hsu, N. C.; Levy, R. C.; Garay, M. J.; Kalashnikova, O. V.; Kahn, R. A.; Brauer, M.; Apte, J. S.; Henze, D. K.; Zhang, L.; Zhang, Q.; Ford, B.; Pierce, J. R.; and Martin, R. V. 2020. Global Estimates and Long-Term Trends of Fine Particulate Matter Concentrations (1998–2018). *Environmental Science & Technology*, 54(13): 7879–7890. Publisher: American Chemical Society.

Harpole, A.; Zingale, M.; Hawke, I.; and Chegini, T. 2019. pyro: a framework for hydrodynamics explorations and prototyping. *Journal of Open Source Software*, 4(34): 1265.

Hawkins, M. D.; Brown, V.; and Ferrell, J. 2017. Assessment of NOAA National Weather Service Methods to Warn for Extreme Heat Events. *Weather, Climate, and Society*, 9(1): 5–13.

Heo, S.; Bell, M. L.; and Lee, J.-T. 2019. Comparison of health risks by heat wave definition: Applicability of wet-bulb globe temperature for heat wave criteria. *Environmental Research*, 168: 158–170.

Heuillet, A.; Couthouis, F.; and Díaz-Rodríguez, N. 2021. Explainability in deep reinforcement learning. *Knowledge-Based Systems*, 214: 106685.

Hondula, D. M.; Meltzer, S.; Balling, R. C.; and Iñiguez, P. 2022. Spatial Analysis of United States National Weather Service Excessive Heat Warnings and Heat Advisories. *Bulletin of the American Meteorological Society*, 103(9): E2017–E2031. Publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society.

Laber, E. B.; Wu, F.; Munera, C.; Lipkovich, I.; Colucci, S.; and Ripa, S. 2018. Identifying optimal dosage regimes under safety constraints: An application to long term opioid treatment of chronic pain. *Statistics in Medicine*, 37(9): 1407–1418. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.7566.

Lee, J. N.; Agarwal, A.; Dann, C.; and Zhang, T. 2023. Learning in POMDPs is Sample-Efficient with Hindsight Observability. ArXiv:2301.13857 [cs, stat].

Levine, A.; Stone, P.; and Zhang, A. 2024. Multistep Inverse Is Not All You Need. *Workshop on Reinforcement Learning Beyond Rewards at the Reinforcement Learning Conference*.

Levine, S.; Kumar, A.; Tucker, G.; and Fu, J. 2020. Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems. *arXiv:2005.01643 [cs, stat]*. ArXiv: 2005.01643.

Li, Y.; Zheng, Y.; and Yang, Q. 2018. Dynamic Bike Reposition: A Spatio-Temporal Reinforcement Learning Approach. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1724–1733. London United Kingdom: ACM. ISBN 978-1-4503-5552-0.

Liao, P.; Greenewald, K.; Klasnja, P.; and Murphy, S. 2020. Personalized HeartSteps: A Reinforcement Learning Algorithm for Optimizing Physical Activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1): 18:1–18:22.

Masselot, P.; Chebana, F.; Campagna, C.; Lavigne, E.; Ouarda, T. B.; and Gosselin, P. 2021. Machine Learning Approaches to Identify Thresholds in a Heat-Health Warning System Context. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184(4): 1326–1346.

Microsoft AI for Good Research Lab. 2021. U.S. Broadband Usage Percentages Dataset.

MIT Election Data and Science Lab. 2018. County Presidential Election Returns 2000-2020.

Mu, T.; Theocharous, G.; Arbour, D.; and Brunskill, E. 2021. Constraint Sampling Reinforcement Learning: Incorporating Expertise For Faster Learning. ArXiv:2112.15221 [cs].

Nahum-Shani, I.; Smith, S. N.; Spring, B. J.; Collins, L. M.; Witkiewitz, K.; Tewari, A.; and Murphy, S. A. 2017. Just-in-Time Adaptive Interventions (JITAIs) in Mobile Health: Key Components and Design Principles for Ongoing Health Behavior Support.

Annals of Behavioral Medicine: A Publication of the Society of Behavioral Medicine, 52(6): 446–462.

Narayanan, S.; Lage, I.; and Doshi-Velez, F. 2022. (When) Are Contrastive Explanations of Reinforcement Learning Helpful? ArXiv:2211.07719 [cs].

Ng, C. F. S.; Ueda, K.; Takeuchi, A.; Nitta, H.; Konishi, S.; Bagrowicz, R.; Watanabe, C.; and Takami, A. 2014. Sociogeographic Variation in the Effects of Heat and Cold on Daily Mortality in Japan. *Journal of Epidemiology*, 24(1): 15–24.

Padakandla, S. 2021. A survey of reinforcement learning algorithms for dynamically varying environments. *ACM Computing Surveys (CSUR)*, 54(6): 1–25.

Pearl, J. 2012. The mediation formula: A guide to the assessment of causal pathways in nonlinear models. *Causality: Statistical perspectives and applications*, 151–179.

Puiutta, E.; and Veith, E. M. 2020. Explainable Reinforcement Learning: A Survey. ArXiv:2005.06247 [cs, stat].

Puterman, M. L. 2014. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Raffin, A.; Hill, A.; Gleave, A.; Kanervisto, A.; Ernestus, M.; and Dormann, N. 2021. Stable-Baselines3: Reliable Reinforcement Learning Implementations. *Journal of Machine Learning Research*, 22(268): 1–8.

Ray, A.; Achiam, J.; and Amodei, D. 2019. Benchmarking Safe Exploration in Deep Reinforcement Learning. Technical report, OpenAI.

Romoff, J.; Henderson, P.; Piche, A.; Francois-Lavet, V.; and Pineau, J. 2018. Reward Estimation for Variance Reduction in Deep Reinforcement Learning. In *Conference on Robot Learning*, 674–699. PMLR.

Sinclair, S. R.; Vieira Frujeri, F.; Cheng, C.-A.; Marshall, L.; Barbalho, H. D. O.; Li, J.; Neville, J.; Menache, I.; and Swaminathan, A. 2023. Hindsight Learning for MDPs with Exogenous Inputs. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 31877–31914. PMLR.

Stuart, N. A.; Hartfield, G.; Schultz, D. M.; Wilson, K.; West, G.; Hoffman, R.; Lackmann, G.; Brooks, H.; Roebber, P.; Bals-Elsholz, T.; Obermeier, H.; Judt, F.; Market, P.; Nietfeld, D.; Telfeyan, B.; DePodwin, D.; Fries, J.; Abrams, E.; and Shields, J. 2022. The Evolving Role of Humans in Weather Prediction and Communication. *Bulletin of the American Meteorological Society*, 103(8): E1720–E1746. Publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement Learning: An Introduction*. The MIT Press, second edition.

Sutton, R. S.; McAllester, D.; Singh, S.; and Mansour, Y. 1999. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12.

Tec, M.; Duan, Y.; and Müller, P. 2023. A Comparative Tutorial of Bayesian Sequential Design and Reinforcement Learning. *The American Statistician*, 77(2): 223–233.

Towers, M.; Terry, J. K.; Kwiatkowski, A.; Balis, J. U.; Cola, G. d.; Deleu, T.; Goulão, M.; Kallinteris, A.; KG, A.; Krimmel, M.; Perez-Vicente, R.; Pierré, A.; Schulhoff, S.; Tai, J. J.; Shen, A. T. J.; and Younis, O. G. 2023. Gymnasium.

Tzikas, D. G.; Likas, A. C.; and Galatsanos, N. P. 2008. The variational approximation for Bayesian inference. *IEEE Signal Processing Magazine*, 25(6): 131–146.

Uehara, M.; Shi, C.; and Kallus, N. 2022. A Review of Off-Policy Evaluation in Reinforcement Learning. ArXiv:2212.06355 [cs, math, stat].

U.S. Census Bureau. 2014. 2009-2013 American Community Survey 5-year County-level Estimates of Population and Median Household Income.

U.S. Energy Information Administration. 2020. Climate Zones - DOE Building America Program.

van der Waa, J.; van Diggelen, J.; Bosch, K. v. d.; and Neerincx, M. 2018. Contrastive Explanations for Reinforcement Learning in terms of Expected Consequences. ArXiv:1807.08706 [cs, stat].

Weinberger, K. R.; Wu, X.; Sun, S.; Spangler, K. R.; Nori-Sarma, A.; Schwartz, J.; Requia, W.; Sabath, B. M.; Braun, D.; Zanobetti, A.; Dominici, F.; and Wellenius, G. A. 2021. Heat warnings, mortality, and hospital admissions among older adults in the United States. *Environment International*, 157: 106834.

Weinberger, K. R.; Zanobetti, A.; Schwartz, J.; and Wellenius, G. A. 2018. Effectiveness of National Weather Service heat alerts in preventing mortality in 20 US cities. *Environment International*, 116: 30–38.

Wu, Q.; Chen, X.; Zhou, Z.; Chen, L.; and Zhang, J. 2021. Deep Reinforcement Learning With Spatio-Temporal Traffic Forecasting for Data-Driven Base Station Sleep Control. *IEEE/ACM Transactions on Networking*, 29(2): 935–948.

Wu, X.; Weinberger, K. R.; Wellenius, G. A.; Dominici, F.; and Braun, D. 2023. Assessing the causal effects of a stochastic intervention in time series data: are heat alerts effective in preventing deaths and hospitalizations? *Biostatistics*.

Xu, H.; Zhan, X.; and Zhu, X. 2022. Constraints Penalized Q-learning for Safe Offline Reinforcement Learning. ArXiv:2107.09003 [cs].

Zajonc, T. 2012. Bayesian Inference for Dynamic Treatment Regimes: Mobility, Equity, and Efficiency in Student Tracking. *Journal of the American Statistical Association*, 107(497): 80–92.

Zanobetti, A.; O'Neill, M. S.; Gronlund, C. J.; and Schwartz, J. D. 2013. Susceptibility to Mortality in Weather Extremes: Effect Modification by Personal and Small Area Characteristics In a Multi-City Case-Only Analysis. *Epidemiology (Cambridge, Mass.)*, 24(6): 809–819.