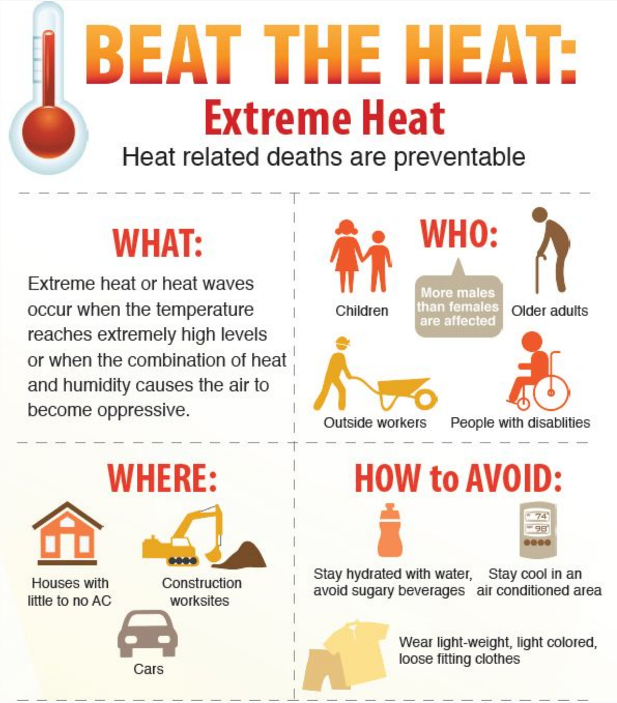
HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

Optimizing Heat Alert Issuance with Reinforcement Learning

Ellen Considine, collaborating with Mauricio Tec
Supervised by Rachel Nethery and Francesca Dominici

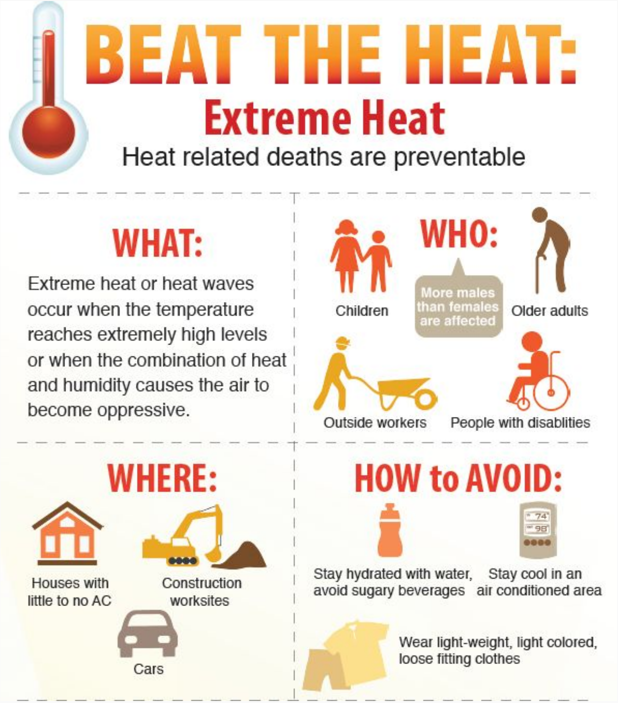
ENAR – March 12, 2024

Extreme heat, public health, & heat alerts



Source: CDC

Extreme heat, public health, & heat alerts



Source: CDC

Context

- Current issuance of heat alerts by the National Weather Service (NWS) does not take advantage of modern data science tools

Context

- Current issuance of heat alerts by the National Weather Service (NWS) does not take advantage of modern data science tools
    - Decision to issue an alert is based on temperature thresholds (differing in northern and southern states)
    - Also strongly affected by the discretion of the local office → stochasticity

Context

- Current issuance of heat alerts by the National Weather Service (NWS) does not take advantage of modern data science tools
    - Decision to issue an alert is based on temperature thresholds (differing in northern and southern states)
    - Also strongly affected by the discretion of the local office → stochasticity
- 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

Context (cont.)

- Past work → some evidence of heat alerts being health-protective at the county level

Context (cont.)

- Past work → some evidence of heat alerts being health-protective at the county level
    - Large uncertainty and substantial heterogeneity

Context (cont.)

- Past work → some evidence of heat alerts being health-protective at the county level
  - Large uncertainty and substantial heterogeneity
- Meanwhile: a growing movement in statistics and public health of focusing on **policy optimization in addition to effect estimation**

Related work on heat alert optimization

Related work on heat alert optimization

1. Wu et al. 2023: developed a causal inference technique for stochastic interventions to infer whether increasing the **probability of issuing** a heat alert would be beneficial

Related work on heat alert optimization

1. Wu et al. 2023: developed a causal inference technique for stochastic interventions to infer whether increasing the **probability of issuing** a heat alert would be beneficial
2. Masselot et al. 2021: compared methods to identify localized **thresholds** above which heat alerts should always be issued

Related work on heat alert optimization

1. Wu et al. 2023: developed a causal inference technique for stochastic interventions to infer whether increasing the **probability of issuing** a heat alert would be beneficial
2. Masselot et al. 2021: compared methods to identify localized **thresholds** above which heat alerts should always be issued

*Neither of these approaches addresses the complications of **sequential dependence***

Related work on heat alert optimization

1. Wu et al. 2023: developed a causal inference technique for stochastic interventions to infer whether increasing the **probability of issuing** a heat alert would be beneficial
2. Masselot et al. 2021: compared methods to identify localized **thresholds** above which heat alerts should always be issued

*Neither of these approaches addresses the complications of **sequential dependence***

- Alert fatigue
- Running out of resources to deploy precautionary measures

Intro to reinforcement learning (RL)

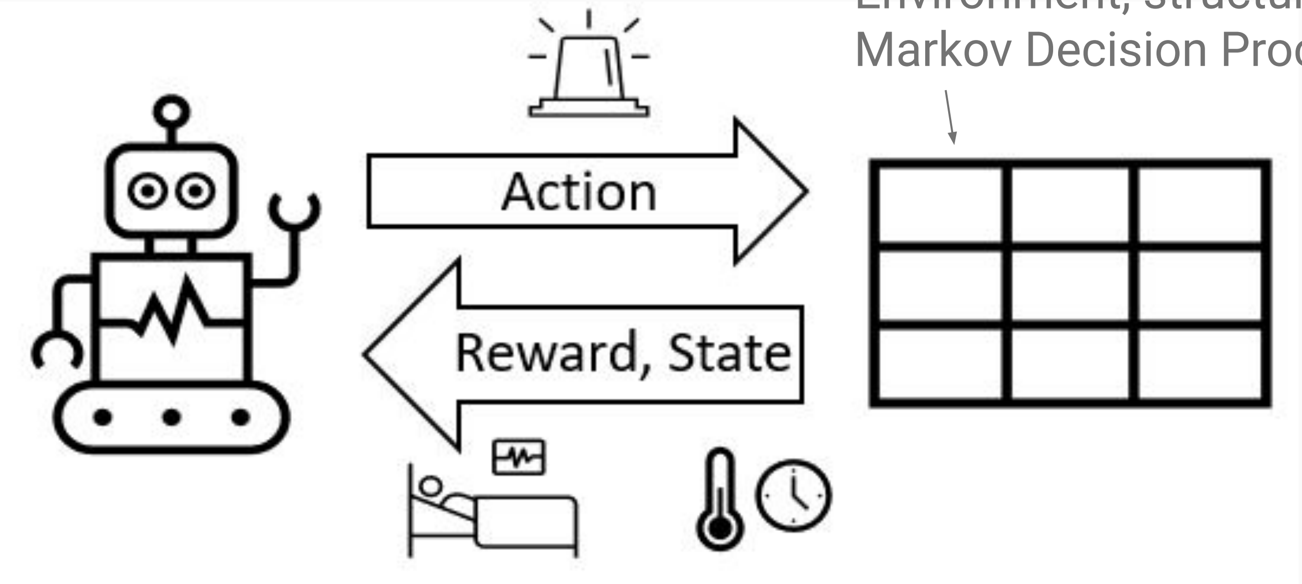
- RL is a rapidly-expanding field of approaches for dealing with sequential decision making (SDM) problems
  - Successful in fields ranging from robotics to mobile health

Intro to reinforcement learning (RL)

- RL is a rapidly-expanding field of approaches for dealing with sequential decision making (SDM) problems
  - Successful in fields ranging from robotics to mobile health
- Algorithmic **agent** uses a **policy** (e.g. when to issue a heat alert) to interact with an **environment**/system to maximize/minimize its **reward**/penalty (e.g. deaths or hospitalizations)

RL in the heat alerts setting

Environment, structured as a Markov Decision Process (MDP)



Heat alerts MDP

Daily, US county-level data → each **episode** is one county-summer (May - Sept.)

Heat alerts MDP

Daily, US county-level data → each **episode** is one county-summer (May - Sept.)

- State:
  - **Exogenous:** *quantile* of heat index or QHI (2006-2016), day of summer, weekend
  - **Endogenous:** recent alert history, remaining alert budget (stay tuned!)

Heat alerts MDP

Daily, US county-level data → each **episode** is one county-summer (May - Sept.)

- State:
  - **Exogenous:** *quantile* of heat index or QHI (2006-2016), day of summer, weekend
  - **Endogenous:** recent alert history, remaining alert budget (stay tuned!)
- Action: issue (1) or do not issue (0) a heat alert

Heat alerts MDP

Daily, US county-level data → each **episode** is one county-summer (May - Sept.)

- State:
  - **Exogenous:** *quantile* of heat index or QHI (2006-2016), day of summer, weekend
  - **Endogenous:** recent alert history, remaining alert budget (stay tuned!)
- Action: issue (1) or do not issue (0) a heat alert
- Reward: rate of heat-related hospitalizations (among Medicare enrollees), transformed such that fewer hospitalizations = greater reward

Major challenges

Major challenges

- **Low signal** (small and easily confounded effects) in observational environmental health datasets
    - Rare events and low signal have been shown to challenge algorithmic decision making

Major challenges

- **Low signal** (small and easily confounded effects) in observational environmental health datasets
  - Rare events and low signal have been shown to challenge algorithmic decision making
- **Spatial variability** in heat alert-health relationship due to geographic self-selection, climate adaptation, socioeconomic status, population density, political ideology, and environmental co-exposures (e.g. AQ)
  - Mainstream RL/SDM methods are not suitable for spatially heterogeneous settings

Major challenges

- **Low signal** (small and easily confounded effects) in observational environmental health datasets
    - Rare events and low signal have been shown to challenge algorithmic decision making
- **Spatial variability** in heat alert-health relationship due to geographic self-selection, climate adaptation, socioeconomic status, population density, political ideology, and environmental co-exposures (e.g. AQ)
    - Mainstream RL/SDM methods are not suitable for spatially heterogeneous settings
- **Limited intervention budget** (esp. to compare with NWS policy)

Novel framework to address challenges

Novel framework to address challenges

1. Create a realistic SDM environment with which to train and evaluate RL and other counterfactual policies relative to the observed NWS policy

Novel framework to address challenges

1. Create a realistic SDM environment with which to train and evaluate RL and other counterfactual policies relative to the observed NWS policy
2. Run single-county RL and provide domain-relevant insights about the results

Novel framework to address challenges

1. Create a realistic SDM environment with which to train and evaluate RL and other counterfactual policies relative to the observed NWS policy
2. Run single-county RL and provide domain-relevant insights about the results
    a. Enables using standard RL algorithms that were not designed for spatially heterogeneous systems
    b. Simplifies the state space by getting to ignore time-fixed covariates during RL training

Novel framework to address challenges

1. Create a realistic SDM environment with which to train and evaluate RL and other counterfactual policies relative to the observed NWS policy
2. Run single-county RL and provide domain-relevant insights about the results

RL environment / simulator

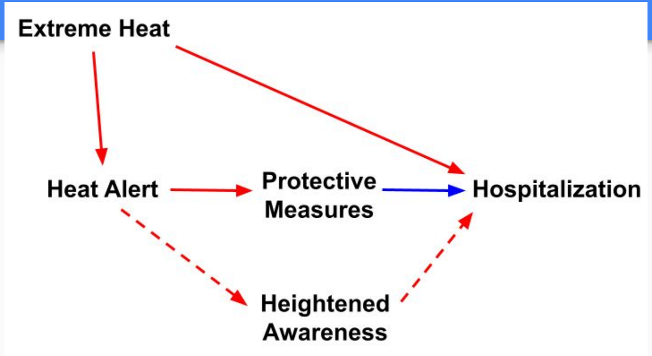
Need:

1. Reward function *R*: generate r_t given (s_t, a_t)
2. Transition function *P*: generate s_{t+1} given (s_t, a_t)

Overview of the rewards model

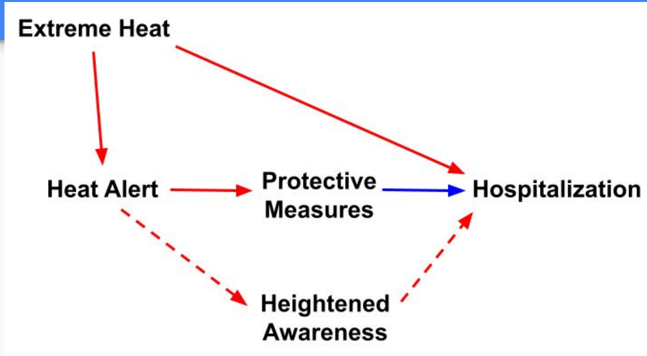
Overview of the rewards model

- Careful specification of the outcome to avoid mediation by heightened awareness of current symptoms



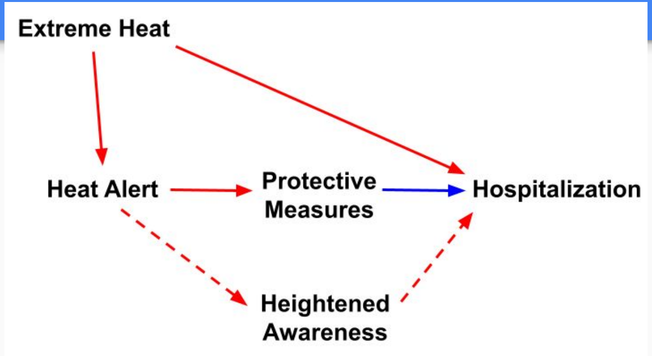
Overview of the rewards model

- Careful specification of the outcome to avoid mediation by heightened awareness of current symptoms



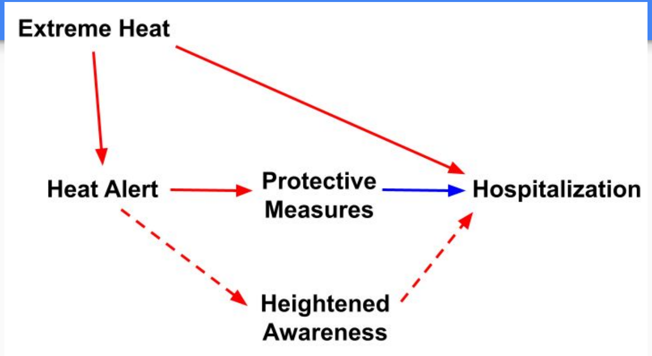
- Bayesian hierarchical model → county-specific coefficients for both the baseline hospitalization rate and the effectiveness of heat alerts

Overview of the rewards model



- Careful specification of the outcome to avoid mediation by heightened awareness of current symptoms
- Bayesian hierarchical model → county-specific coefficients for both the baseline hospitalization rate and the effectiveness of heat alerts
- Data-driven prior using spatial features

Overview of the rewards model



- Careful specification of the outcome to avoid mediation by heightened awareness of current symptoms
- Bayesian hierarchical model → county-specific coefficients for both the baseline hospitalization rate and the effectiveness of heat alerts
- Data-driven prior using spatial features
- Variational inference to handle the high dimensionality of parameters

Overview of the transition model

$P(s_{t+1}|s_t, a_t)$

Overview of the transition model

$$P(s_{t+1}|s_t, a_t) = P((\xi_{t+1}, x_{t+1})|(\xi_t, x_t), a_t) = P_\xi(\xi_{t+1}|\xi_t)P_x(x_{t+1}|x_t, a_t)$$

where ξ is exogenous and *x* is endogenous

Overview of the transition model

$$P(s_{t+1}|s_t, a_t) = P((\xi_{t+1}, x_{t+1})|(\xi_t, x_t), a_t) = P_\xi(\xi_{t+1}|\xi_t)P_x(x_{t+1}|x_t, a_t)$$

where ξ is exogenous and *x* is endogenous

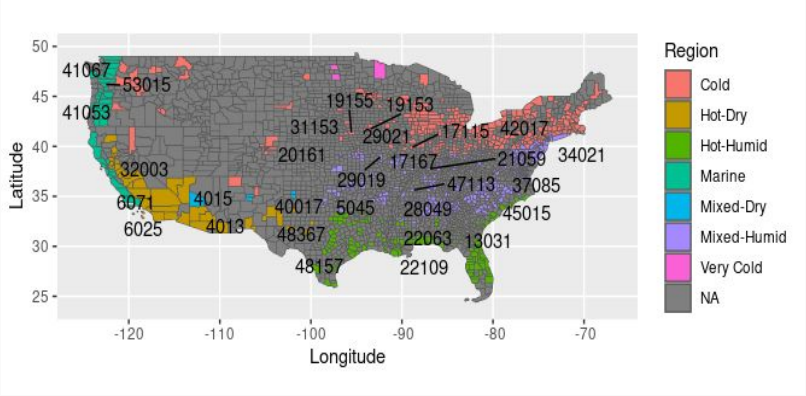
→ Rather than introduce error via modeling, *sample* weather trajectories!

Overview of the transition model

$$P(s_{t+1}|s_t, a_t) = P((\xi_{t+1}, x_{t+1})|(\xi_t, x_t), a_t) = P_\xi(\xi_{t+1}|\xi_t)P_x(x_{t+1}|x_t, a_t)$$

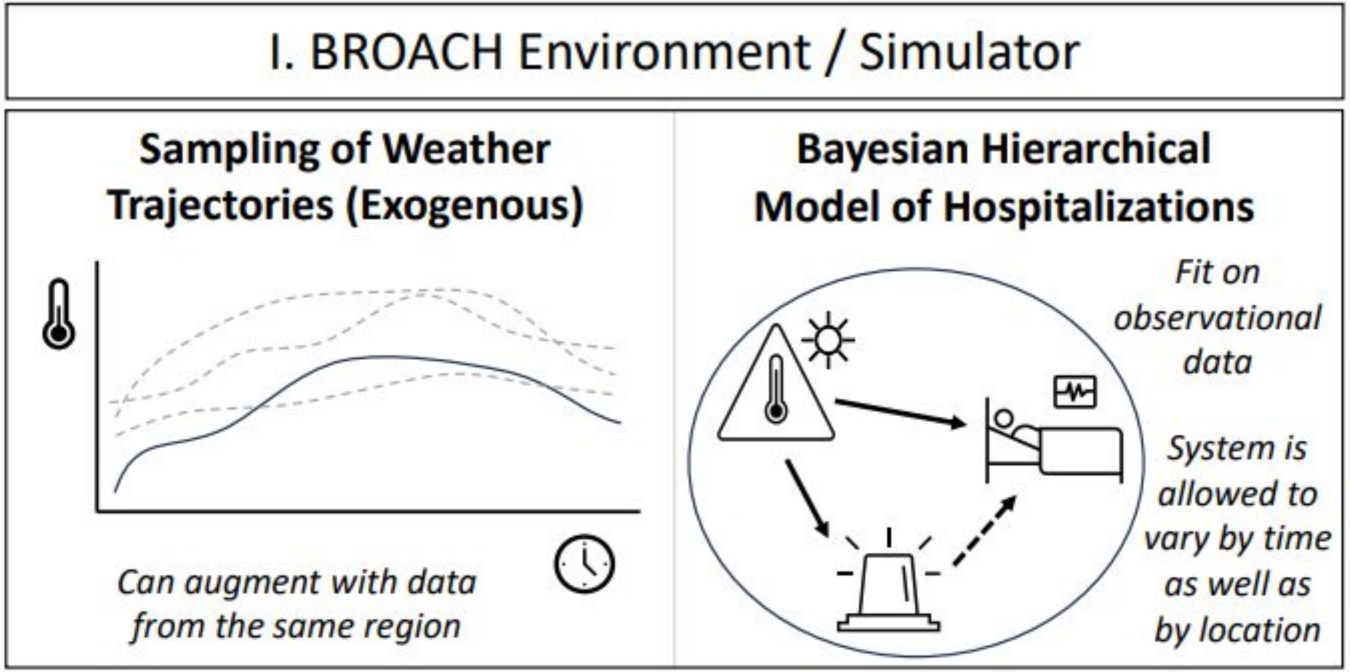
where ξ is exogenous and *x* is endogenous

→ Rather than introduce error via modeling, *sample* weather trajectories!



To avoid overfitting 2006-2016 during single-county RL, *augment* the exogenous data with weather trajectories from other counties in the same regional climate zone

Recap: Bayesian Rewards Over Actual Climate History



Novel framework to address challenges

1. Create a realistic SDM environment with which to train and evaluate RL and other counterfactual policies relative to the observed NWS policy
2. Run single-county RL and provide domain-relevant insights about the results

30,000 foot view of RL algorithms

- Three major families of algorithms: value (Q) learning, policy learning, and actor-critic

30,000 foot view of RL algorithms

- Three major families of algorithms: value (Q) learning, policy learning, and actor-critic
- Important differentiator: how is *exploration* induced?
  - Deterministic policy with epsilon-greedy (choose action at random with prob ε)
  - Stochastic policy (the policy itself is a probability distribution → sample from it)

30,000 foot view of RL algorithms

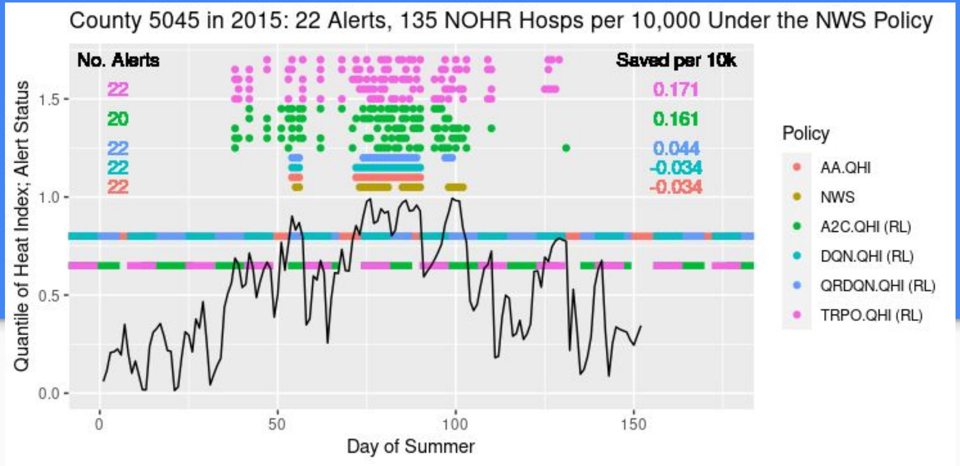
- Three major families of algorithms: value (Q) learning, policy learning, and actor-critic
- Important differentiator: how is *exploration* induced?
  - Deterministic policy with epsilon-greedy (choose action at random with prob ε)
  - Stochastic policy (the policy itself is a probability distribution → sample from it)
- We investigate using four of the most widely used RL algorithms
  - Both deterministic and stochastic
  - Including Q-learning, policy learning, and actor-critic

Constrained RL

Two related problems:

- Can’t issue too many alerts
- Heat alerts only make sense on very hot days

Constrained RL



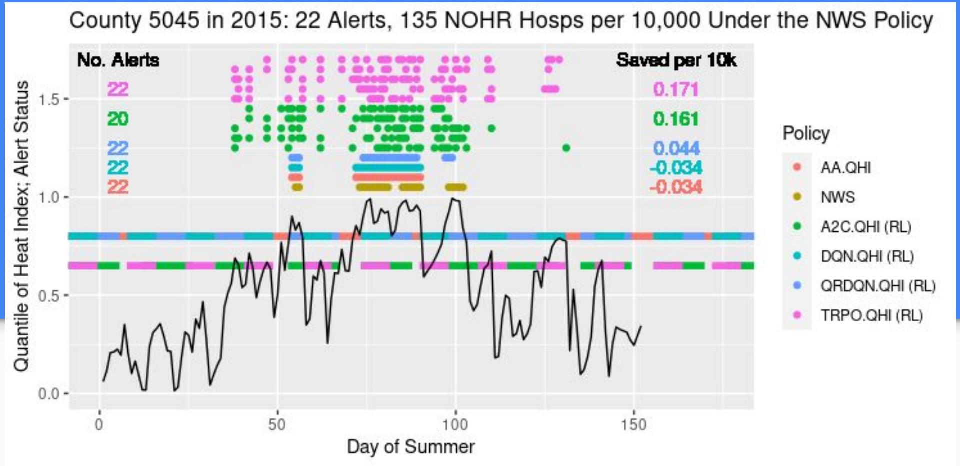
Two related problems:

- Can't issue too many alerts
- Heat alerts only make sense on very hot days

Our approach:

- Strictly limit the number of alerts in an episode to that issued by the NWS
  - Allows exact comparison with NWS policy using modeled rewards

Constrained RL



Two related problems:

- Can't issue too many alerts
- Heat alerts only make sense on very hot days

Our approach:

- Strictly limit the number of alerts in an episode to that issued by the NWS
  - Allows exact comparison with NWS policy using modeled rewards
- Restrict issuance of alerts to very hot days (suffix ".QHI")
  - Identify optimal threshold for each county

Main RL results

- Standard RL methods perform worse than NWS

Main RL results

- Standard RL methods perform worse than NWS
- RL policies modified by our QHI restriction outperform the NWS with statistical significance

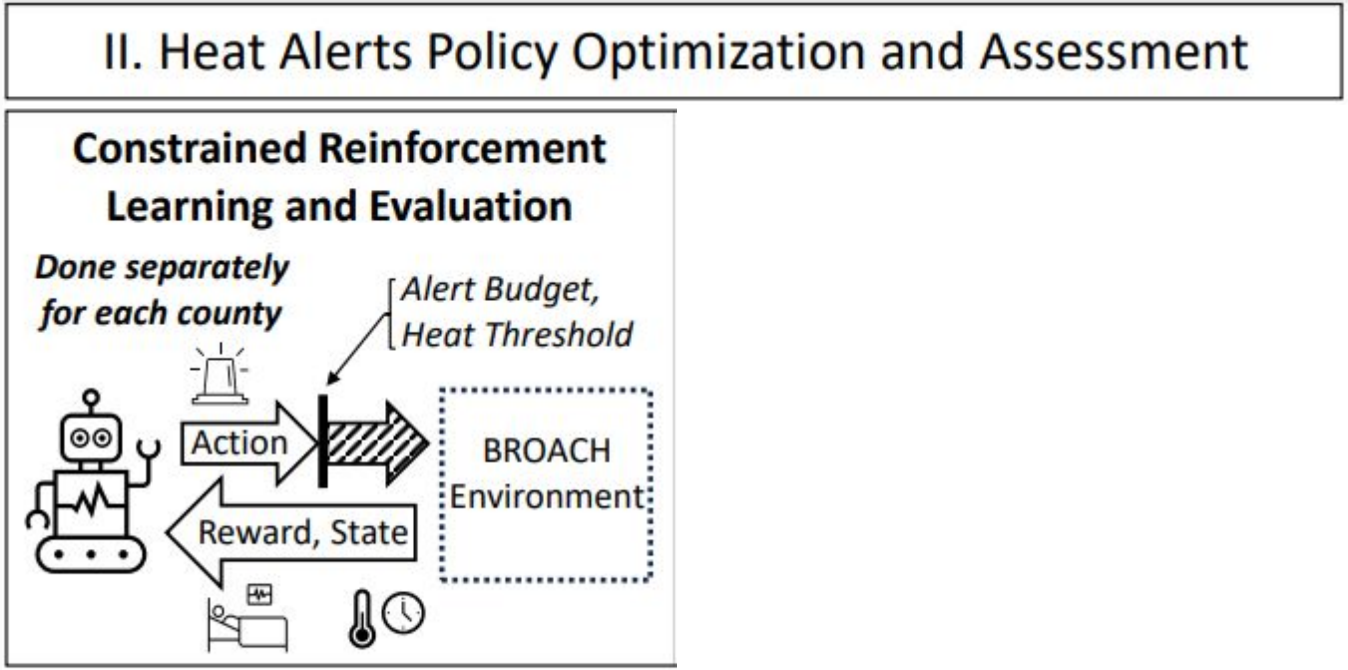
Main RL results

- Standard RL methods perform worse than NWS
- RL policies modified by our QHI restriction outperform the NWS with statistical significance
    - But there was large heterogeneity across counties

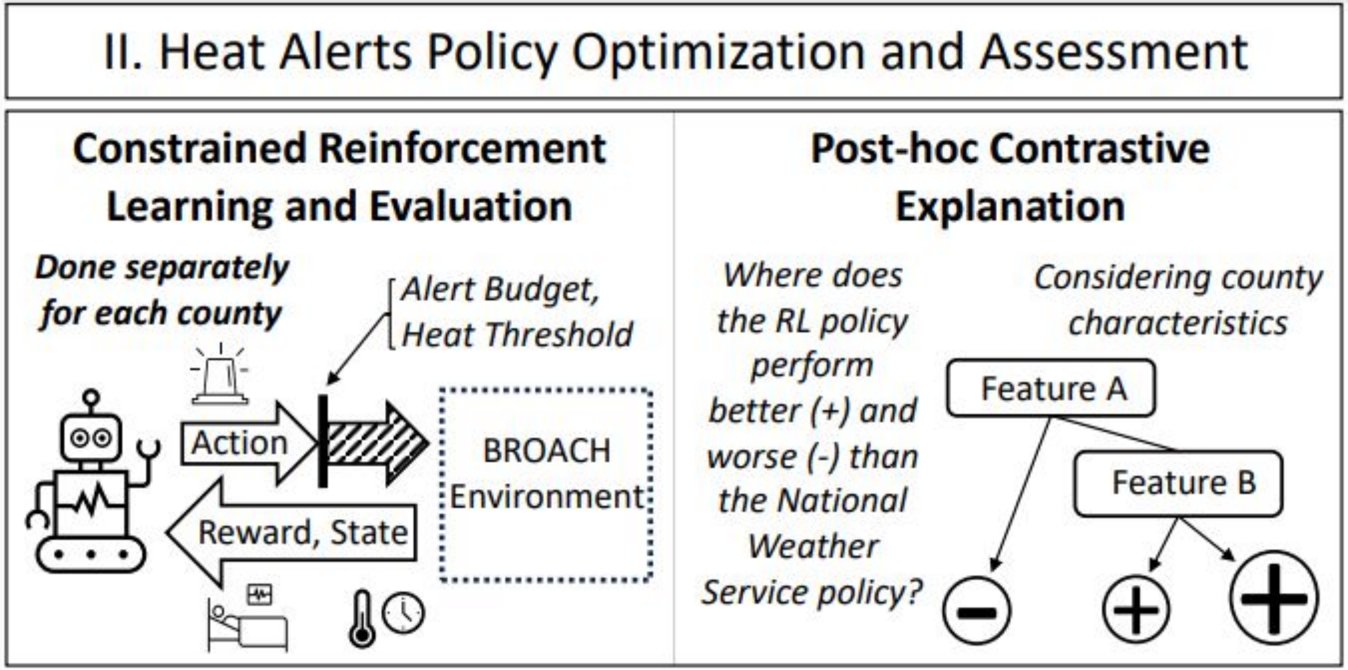
Main RL results

- Standard RL methods perform worse than NWS
- RL policies modified by our QHI restriction outperform the NWS with statistical significance
  - But there was large heterogeneity across counties
- Best RL issued **stochastic** policies

Recap: part 2


II. Heat Alerts Policy Optimization and Assessment
Constrained Reinforcement Learning and Evaluation
Done separately for each county
Alert Budget,
Heat Threshold
Action
Reward, State
BROACH Environment

Recap: part 2



II. Heat Alerts Policy Optimization and Assessment
Constrained Reinforcement Learning and Evaluation
Done separately for each county
Alert Budget,
Heat Threshold
Action
BROACH Environment
Reward, State
Post-hoc Contrastive Explanation
Where does the RL policy perform better (+) and worse (-) than the National Weather Service policy?
Considering county characteristics
Feature A
Feature B

CART results

RL performs best in counties with:

CART results

RL performs best in counties with:

- High alert-health signal (as estimated by our rewards model)
  - Especially when it is optimal to issue alerts earlier in the summer than NWS

CART results

RL performs best in counties with:

- High alert-health signal (as estimated by our rewards model)
  - Especially when it is optimal to issue alerts earlier in the summer than NWS
- More prolonged heat waves (indicated by longer streaks of alerts in threshold-based policies such as NWS)

Directions for future work

Directions for future work

- Ensure new policy is never worse than existing policy in each county

Directions for future work

- Ensure new policy is never worse than existing policy in each county
- Determine alert budget in a changing system (e.g. due to climate change)

Directions for future work

- Ensure new policy is never worse than existing policy in each county
- Determine alert budget in a changing system (e.g. due to climate change)
- Multi-objective RL to consider multiple kinds of health data, across age groups

Directions for future work

- Ensure new policy is never worse than existing policy in each county
- Determine alert budget in a changing system (e.g. due to climate change)
- Multi-objective RL to consider multiple kinds of health data, across age groups
- For RL methodologists: development of new general-purpose algorithms that perform robustly in this kind of setting
  - We will publish the BROACH simulator to facilitate

Check out our preprint on arXiv!

Thank you!

Co-authors: Mauricio Tec, Rachel Nethery, Francesca Dominici, and Greg Wellenius

National Studies on Air Pollution & Health (NSAPH) Research Group

National Science Foundation

Additional Slides

Basics of a Markov Decision Process

$$\langle S, A, R, P, \gamma \rangle$$

MDP is a tuple

$$R : S \times A \to \mathbb{R}$$

Expected reward function

$$P : S \times A \times S \to [0, 1]$$

Transition function

$$\pi^*(a_t | s_t) \to [0, 1]$$

Policy

$$J(\pi) = E_\pi \left[\sum_{t=0}^{H-1} \gamma^t R(s_t, a_t) \right]$$

Objective: finite-horizon value function

Discussion

- Modest absolute public health benefit, but cost-effective intervention
  - About 222 NOHR hospitalizations per year saved across US (approximate 95% CI = (-491, 1,131)), using Medicare population from 2011
  - Increases to 262 if under a safe policy s.t. counties which would not benefit are unaffected
  - Also: both frequency of extreme heat events and size of Medicare population are projected to continue increasing
- Palatability of a stochastic policy?
  - Less immediately satisfying
  - But for human-in-the-loop, an algorithm reporting probabilities is more informative
  - In any case, would need to utilize exploration to update an online RL over time

Bayesian hierarchical model for rewards

Let (*j*, *k*) index a county-summer and *t* index time (days in the summer)

$$y_t^{(k,j)}(a) \sim \text{Poisson}(n^{(k,j)} \rho_t^{(k,j)}(a)),$$
$$\rho_t^{(k,j)}(a) := \lambda_k(s_t^{(k,j)})(1 - a \cdot \tau_k(v_t^{(k,j)}))$$

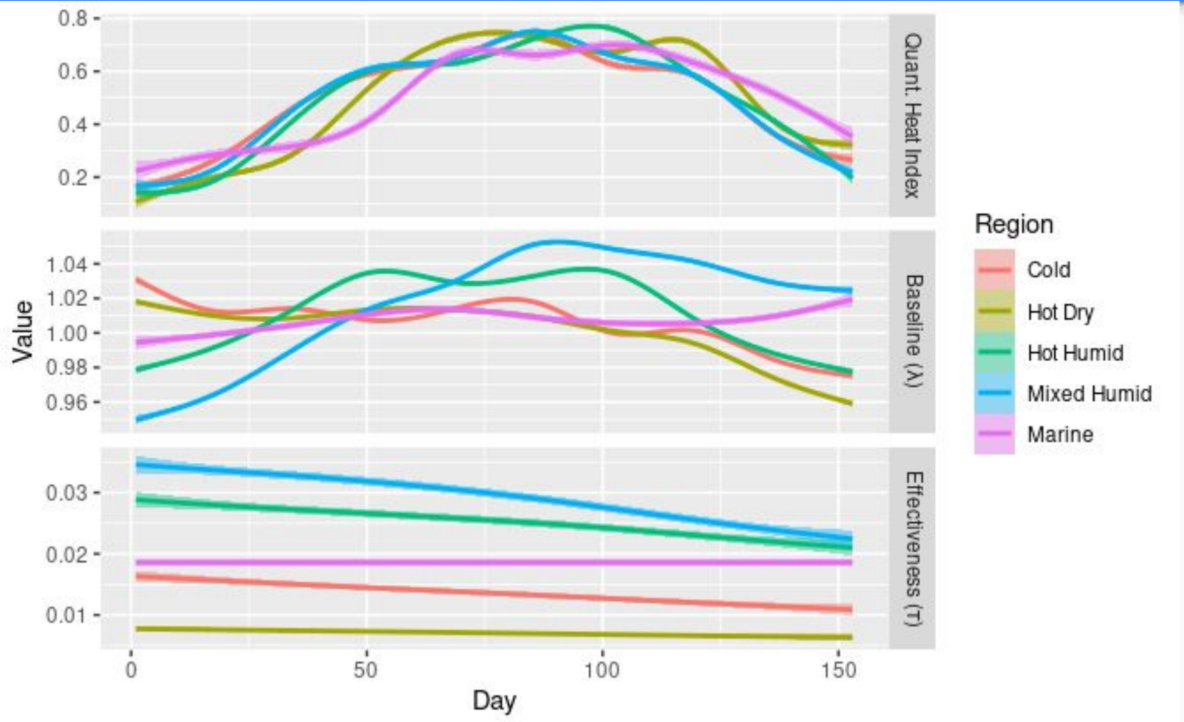
Baseline rate $\lambda(s_t^{(k,j)}) := \exp\left(\beta_k^\top s_t^{(k,j)}\right)$

Alert effectiveness $\tau(v_t^{(k,j)}) := \text{sigmoid}\left(\delta_k^\top v_t^{(k,j)}\right)$

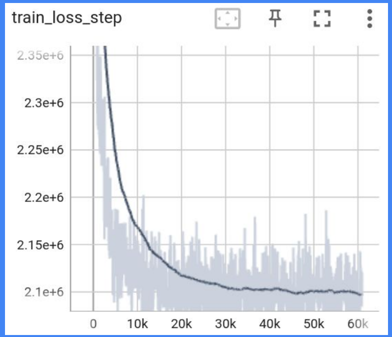
Allows non-linearity through s_t and v_t

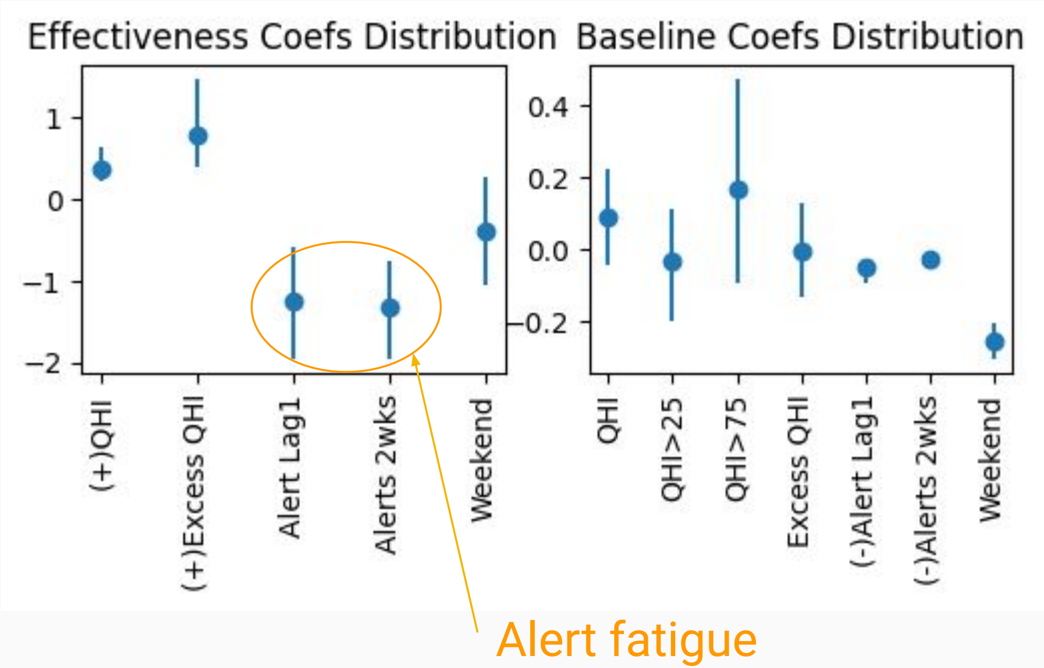
$$\Rightarrow \quad r_t^{(k,j)}(a) := 1 - \rho_t^{(k,j)}(a)$$

Aggregated by regional climate zone



Rewards model results





Displayed very good coverage when we ran it on synthetic data (1,000 samples from the posterior predictive) using known coefficients: average coverage across parameters for 90% CI was 0.897

Rewards model estimation

Only have 11 summers per county… To address this plus low signal:

1. Borrow statistical strength across counties using a data-driven random effects prior (based on spatial features *w*)
2. Inject domain knowledge / assumptions on the sign of certain coefficients

Can be seen as a form of Empirical Bayes

Rewards model estimation

Only have 11 summers per county… To address this plus low-signal:

1. Borrow statistical strength across counties using a data-driven random effects prior (based on spatial features *w*)
2. Inject domain knowledge on the sign of certain coefficients

$$\gamma_k = [\beta_k; \delta_k] = (\gamma_k^\ell)_{\ell=1}^L$$

$$\gamma_k^\ell \sim p(\gamma_k^\ell | \sigma_\ell; \theta_\ell, w_k) = \begin{cases} \text{Normal}(f_{\theta_\ell}(w_k), \sigma_\ell^2) & \text{no domain knowledge}, \\ \text{LogNormal}(\exp(f_{\theta_\ell}(w_k)), \sigma_\ell^2) & \text{if } \gamma_k^\ell \in (0, \infty), \\ \text{NegLogNormal}(-\exp(f_{\theta_\ell}(w_k)), \sigma_\ell^2) & \text{if } \gamma_k^\ell \in (-\infty, 0), \\ \text{Identically zero} & \text{if } \gamma_k^\ell = 0, \end{cases}$$

$$\sigma_\ell \sim \text{HalfCauchy}(0, 1) \qquad \text{where } f_{\theta_\ell} \text{ is a feed-forward neural network with weights } \theta_\ell.$$

Additional details

Domain knowledge-based constraints:

1. Past heat alerts cannot increase the baseline hospitalization rate
2. Higher QHI cannot decrease the effectiveness of heat alerts — note that this is conditional on day of summer

Additional details

Domain knowledge-based constraints:

1. Past heat alerts cannot increase the baseline hospitalization rate
2. Higher QHI cannot decrease the effectiveness of heat alerts — note that this is conditional on day of summer

Model fitting with **Pyro**:

- Use variational inference to handle the high dimensionality of parameters (approximate the true posterior distribution)

Experimental Setup

Ran experiments on 30 counties (spread across the five main climate regions)

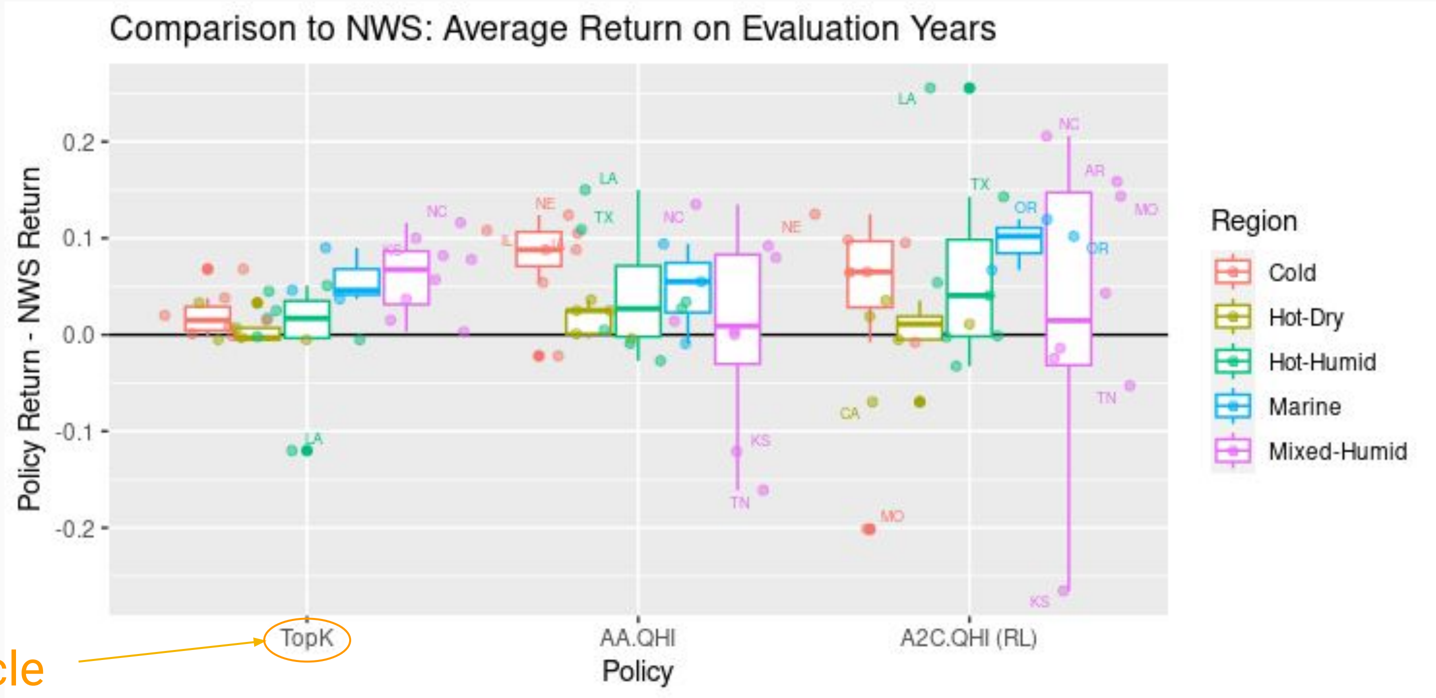
Baselines: other alternative policies

- Random, basic NWS thresholds, top k hottest days, always alert above an optimized threshold

Held-out test years: {2007, 2011, 2015}

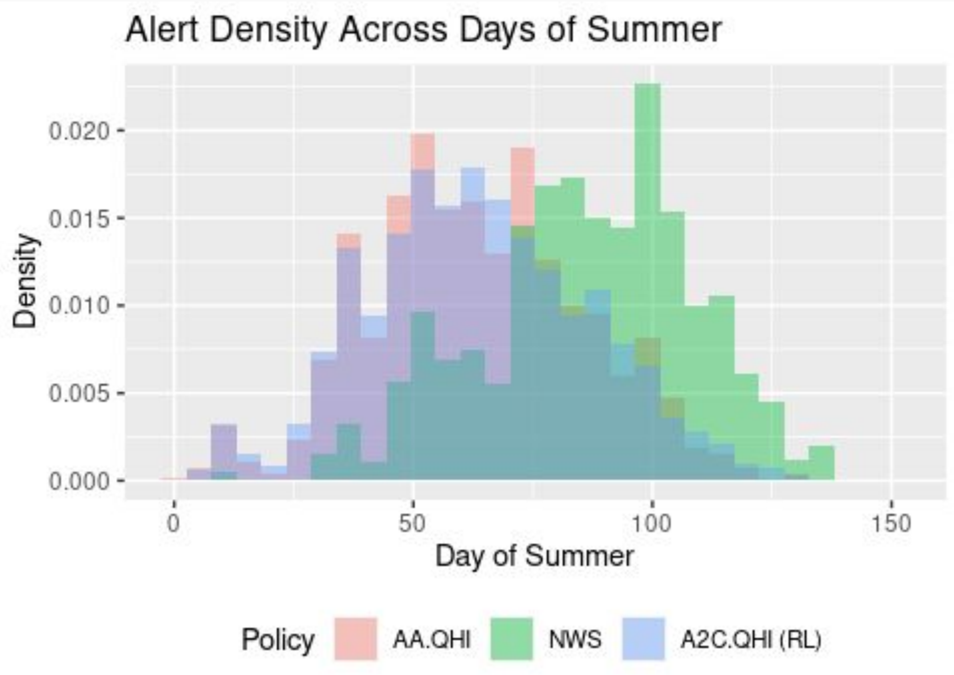
- Training: all counties, training years
- Validation / tuning: all counties *except* the county of interest, testing years
- Final evaluation: county of interest, testing years

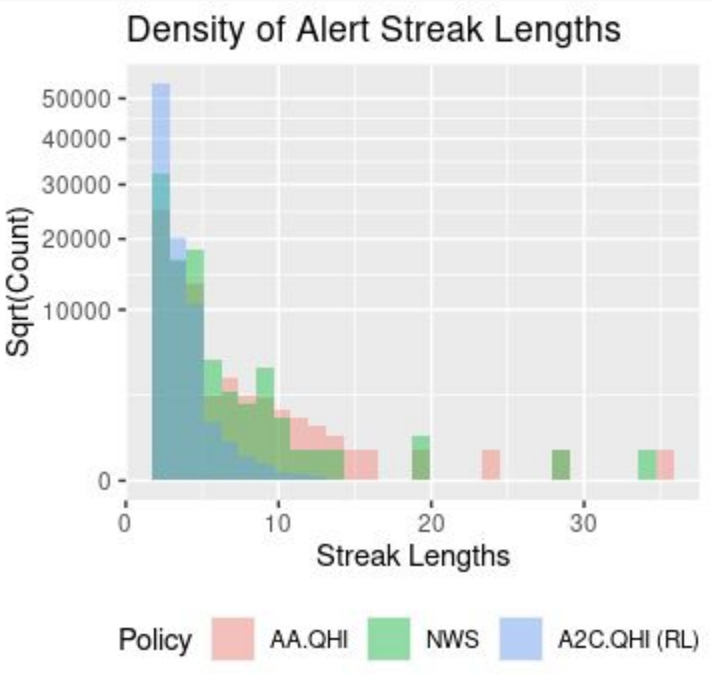
Heterogeneity across counties



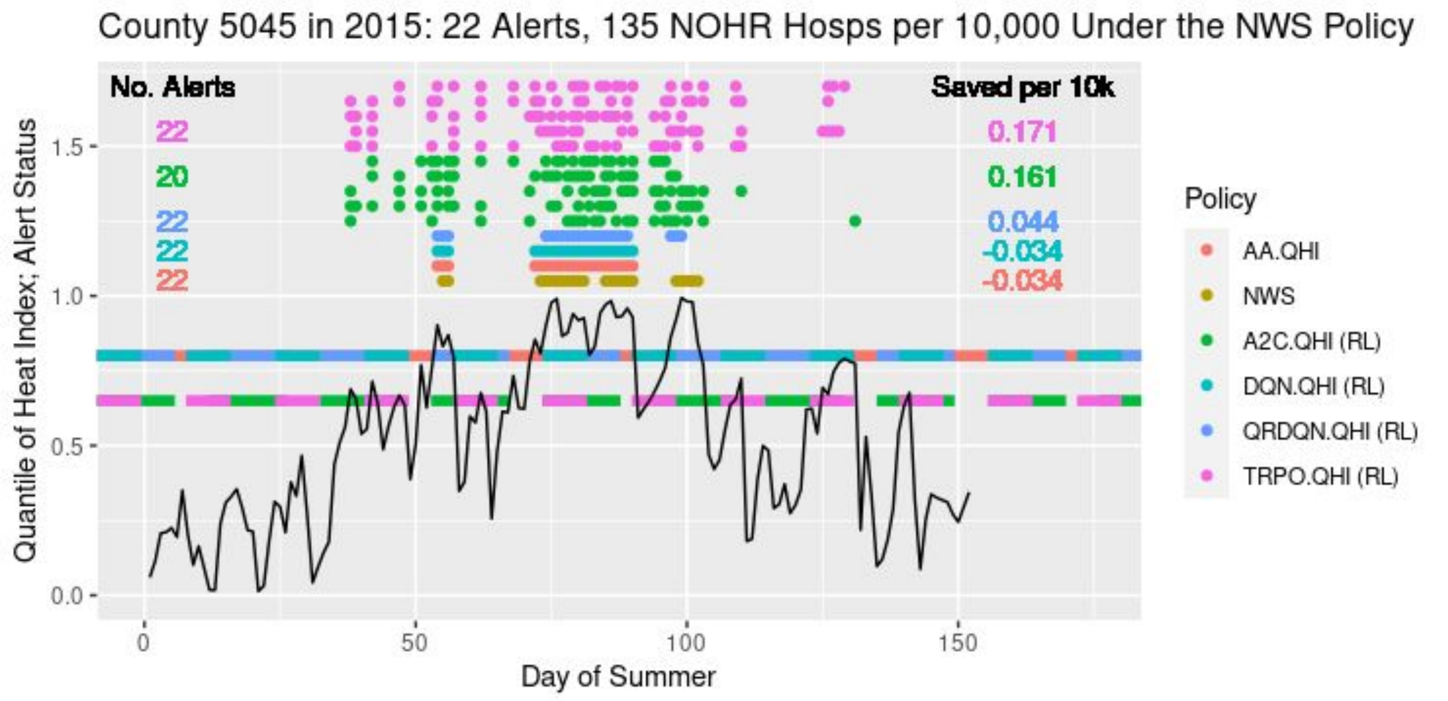
Oracle

Temporal characteristics



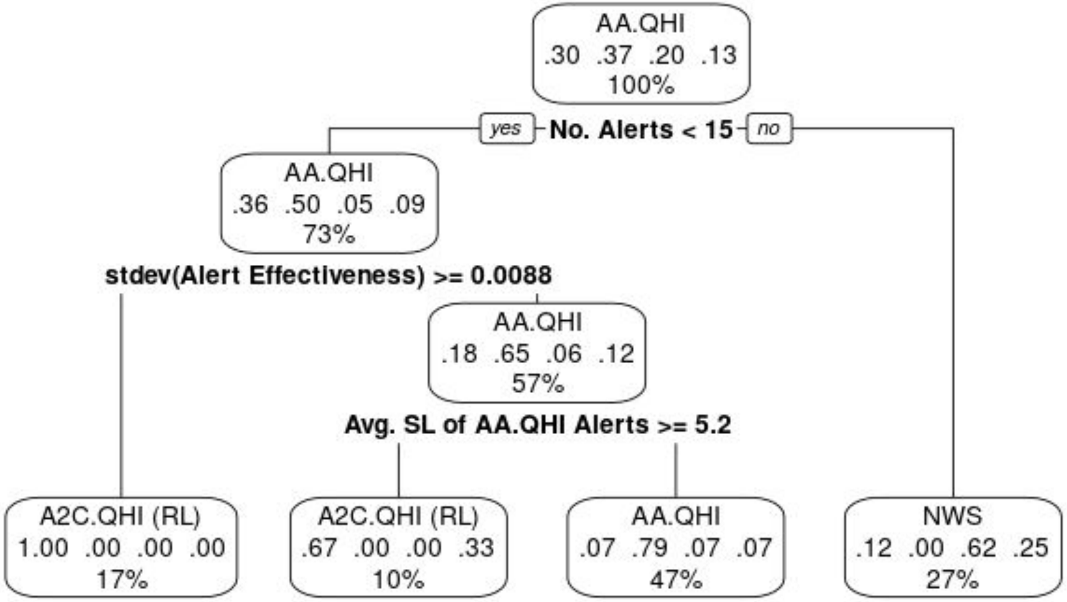


Example county-summer



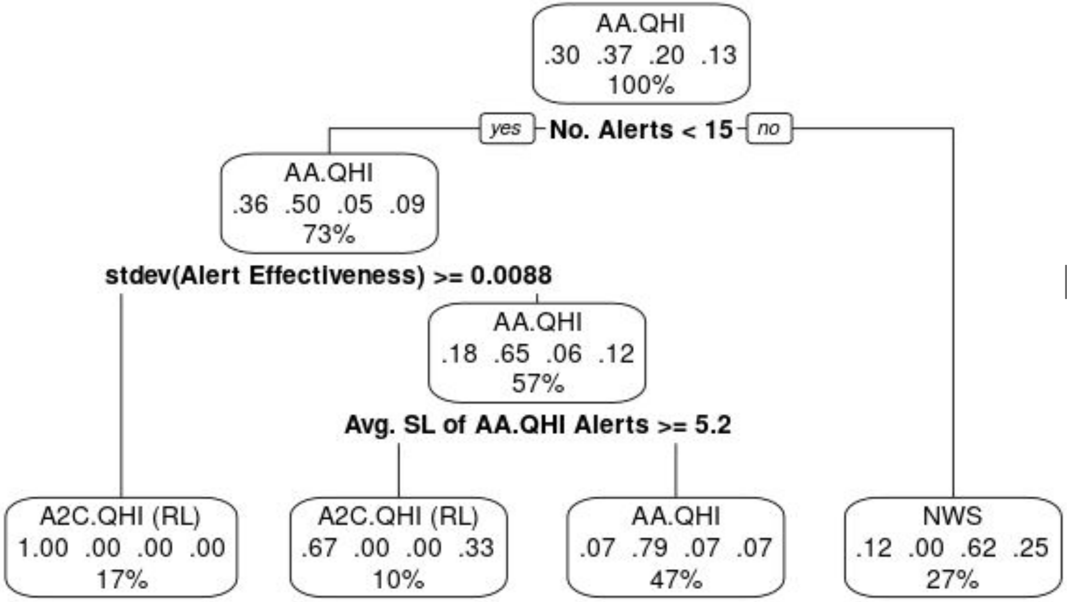
CART: contrastive analysis

Best Policy Type
[Classification probabilities for:
A2C.QHI, AA.QHI, NWS, TRPO.QHI]
Fraction of Counties



CART: contrastive analysis

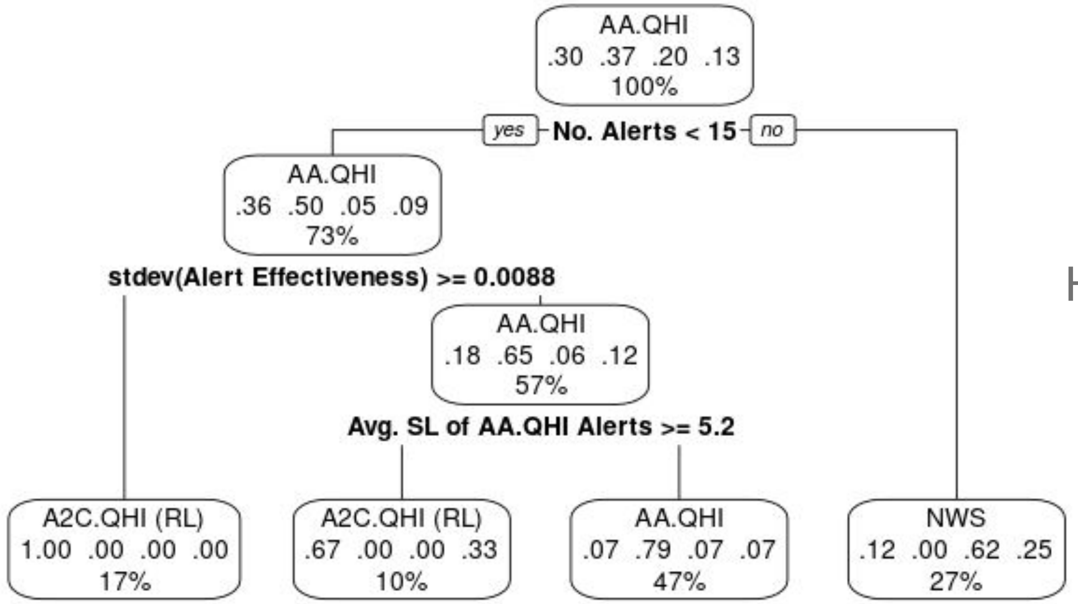
Best Policy Type
[Classification probabilities for:
A2C.QHI, AA.QHI, NWS, TRPO.QHI]
Fraction of Counties



Higher signal ~ higher median HH income

CART: contrastive analysis

Best Policy Type
[Classification probabilities for:
A2C.QHI, AA.QHI, NWS, TRPO.QHI]
Fraction of Counties

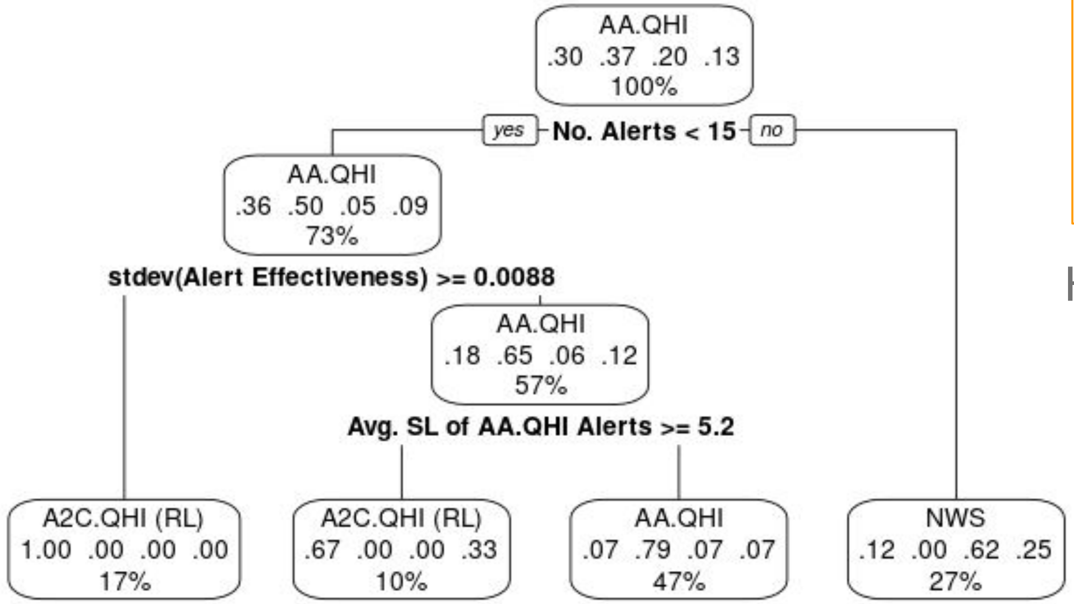


Higher signal ~ higher median HH income

Longer streaks ~ more humid regions

CART: contrastive analysis

Best Policy Type
[Classification probabilities for:
A2C.QHI, AA.QHI, NWS, TRPO.QHI]
Fraction of Counties



Regression tree: RL performs better than NWS when RL determines it is optimal to issue alerts earlier in the summer

Higher signal ~ higher median HH income

Longer streaks ~ more humid regions