Interventionally Consistent Surrogates for Urban Flood Resilience Planning

Francisco Madaleno[1,2], Oskar Bohn Lassen[1], Serio Agriesti[1], Miguel Costa[1], Carlos Lima Azevedo[1], Filipe Rodrigues[1], and Francisco Camara Pereira[1]

[1]Machine Learning for Smart Mobility Lab, Department of Technology, Management and Economics, Danish Technical University, Denmark
[2]fmfsa@dtu.dk

Short summary

Policy evaluation for urban flood resilience requires high-fidelity simulations of interdependent flood and transportation dynamics, but the computational cost of these models—the "Simulation Bottleneck"—prevents their use for scenario analysis and policy optimization. Standard machine learning surrogates offer the needed speed but are typically trained on observational data and fail in interventional scenarios, i.e. they are not *interventionally consistent*. This paper uses an Interventionally Consistent Surrogate framework grounded in causal abstraction theory. Using Batched Multi-Task Gaussian Processes structured to respect the simulator's causal topology, we train a surrogate that approximately preserves the simulator's response under policy interventions. We demonstrate the framework on Copenhagen's inner city using a flood-transport simulator, achieving $R2 > 0.92$ across all three prediction tasks while providing calibrated uncertainty estimates.
Keywords: Causal Abstraction, Gaussian Processes, Surrogate Modelling, Resilience Planning, Structural Causal Models.

1 Introduction

Urban resilience planning involves mitigating the impact of extreme weather events on critical infrastructure. Iterative planning methods—including scenario analysis, sensitivity studies, and policy optimization via Reinforcement Learning (Tang et al., 2025; Ge et al., 2014; Costa et al., 2026)—require an enormous number of simulator evaluations. In high-fidelity settings such as the MAAT framework (Costa et al., 2026), a single forward step can take minutes, making iterative evaluation computationally intractable. Even outside Reinforcement Learning applications, impact evaluation through simulations may be realistically carried over only a limited set of scenarios that does not represent the full spectrum of possibilities due to computational constraints (Osorio & Bierlaire, 2013; Natterer et al., 2025).

Surrogates—machine-learning models mapping input-output relationships from data—are a standard solution, as they are faster than simulators. Yet most surrogates are black-box models trained on observational data and often fail to satisfy *interventional consistency*: the requirement that the surrogate behaves similarly to the simulator under external policy interventions (Dyer et al., 2024). Without this property, a planner may base decisions on statistical artifacts that do not reflect the physical system. We address this by formalizing the surrogate as a *causal abstraction* of the high-dimensional simulator, ensuring that the surrogate's error remains low across the intervention set of interest.

We frame surrogate modeling through Structural Causal Models (SCMs) (Pearl, 2009; Peters et al., 2017) and build upon the theory of *causal abstraction* (Rubenstein et al., 2017; Beckers & Halpern, 2019), recently expanded via cluster DAGs (Anand et al., 2023; Madaleno et al., 2026) and applied to Reinforcement Learning (Kekić et al., 2025). Dyer et al. (2024) formalized the mathematical requirements for a surrogate to achieve strict *interventional consistency*. While theoretically rigorous, scaling these methods to real-world physical systems remains an open challenge. We bridge this gap by translating the theory into a practical, scalable architecture for urban resilience planning.

Our contributions are:

- **Formalism:** We frame the compression of a city-scale flood and traffic simulator (MAAT) into a surrogate as a causal abstraction, ensuring that interventions do not result in unacceptable approximation errors.
- **Method:** We propose an Interventionally Consistent Surrogate (ICS) using Batched Multi-Task Gaussian Processes that capture the underlying SCM, enabling scalable, zone-level prediction with calibrated uncertainty.
- **Empirical Validation:** We demonstrate the approach on Copenhagen's inner city, achieving high predictive accuracy ($R^2 > 0.92$) on multiple indicators across a test set of interventions.

2 The Fine-Grained System: MAAT

We use the MAAT framework (Costa et al., 2026) as our case study simulator. MAAT integrates flood modeling, flood-affected transport simulation, and policy testing of intervention portfolios, and has been applied to RL-based climate adaptation planning in Copenhagen.

In short, this integrated assessment model framework is composed of 3 main modules: 1) a rainfall projection model, 2) a flood model, and 3) a transport simulation model. At each timestep, a rainfall event is sampled following a rainfall projection distribution based on the Danish Climate Institute's Klimaatlas (Danmarks Meteorologiske Institut, 2023) (a model that projects how daily rainfall intensities will increase in time until 2100). After sampling one rainfall event, a flood model maps how water is distributed across a city. Here, we use SCALGO Live (SCALGO Live, Accessed:2024-06-07), which simulates how precipitation water flows depending on the precipitation intensity and terrain data. This allows for water depths to be mapped to roads, cycle paths, and sidewalks across our case study. This water, in turn, affects transport, creating delays and untraversable roads and paths when reaching critical heights. To model how water depths affect travel speed the framework uses depth-disruption functions (Pregnolato et al., 2017) for car, bicycle, and walking trips. Next, trips by car, bicycle, and on foot are simulated for each flood event following real world demand (Christiansen & Baescu, 2024, Transportvaneundersøgelsen). Finally, direct and indirect transport impacts (travel delays, trip cancellations, infrastructure damage) are computed and aggregated over 29 traffic assignment zones (TAZs, each covering a few city blocks), corresponding to those used in the Danish National Transport Model (Vejdirektoratet, 2022, Grøn Mobilitetsmodel).

As a whole, the MAAT framework can be described as a tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \Omega \rangle$, operating over discrete time steps t:

- **State ($\mathcal{S}$):** The full physical state s_t: water depths at every node and link in the transport network graph, coupled with rainfall intensity ($|E| \approx 10^4$ edges, with $|E|$ being the number of edges across the network).
- **Actions ($\mathcal{A}$):** Allowable infrastructure interventions a_t (e.g. increasing drainage in a given location, implementing permeable pavements, etc.)
- **Transition Dynamics ($\mathcal{T}$):**

$$s_{t+1} = f(s_t, a_t, \text{Rain}_t) \tag{1}$$

  Computing f requires solving water flood equations and transport simulation, making time steps in MAAT $\mathcal{T}$ computationally expensive.

- **Outcome (Ω):** Metrics y_t: per-vehicle travel times, trip cancellations, and infrastructure damage costs.

Due to the computational cost of running each time step, evaluating flood control strategies through a high number of runs with MAAT is intractable. Our goal is to compress the simulated system $\mathcal{M}$ into a surrogate that is computationally efficient yet causally consistent with the underlying physics, guaranteeing that each modification to the environment is faithfully captured by the surrogate as well.

3 The Causal Abstraction Framework

We frame surrogate learning as finding a *constructive causal abstraction* of $\mathcal{M}$ (Dyer et al., 2024). Rather than fitting a regression model to observed correlations, we define a coarse-grained SCM $\mathcal{M}'$ that captures the salient cause-effect relationships at the level of Traffic Analysis Zones.
Let the city be partitioned into Z disjoint zones and let M denote the number of transport modes. We define abstraction mappings (τ, ω) that project the high-dimensional variables of $\mathcal{M}$ onto a lower-dimensional space:

1. **State Abstraction (τ_S):** $s_t \mapsto \mathbf{S}_t = [S_{1,t}, \ldots, S_{Z,t}]^\top$, where $S_{z,t} \in \mathbb{R}_{\geq 0}$ is the aggregate flood severity in zone z.

2. **Action Abstraction (ω):** $a_t \mapsto \mathbf{A}'_t = [A'_{1,t}, \ldots, A'_{Z,t}]^\top$, with $A'_{z,t} \in \{0, \ldots, K-1\}$ a categorical flood control decision per zone (K intervention types including the null action).

3. **Outcome Abstraction (τ_Y):** $y_t \mapsto (\mathbf{L}_t, \mathbf{C}_t, \mathbf{I}_t)$, aggregating the outcomes into zone-level metrics for Traffic Delay, Trip Cancellations, and Infrastructure Cost.

The abstraction maps (τ, ω), with $\tau = (\tau_S, \tau_Y)$, reduce the state space to a tractable dimension and define the causal topology that constrains the surrogate architecture.
A surrogate achieves interventional consistency if the diagram in Figure 1 commutes: intervening in the simulator and then abstracting the outcome is equivalent to abstracting the intervention first and running the surrogate.

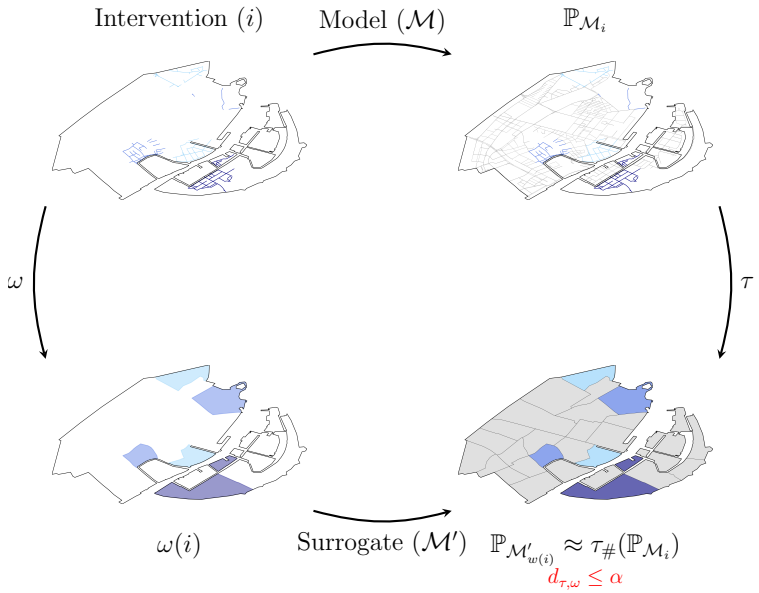


Figure 1: Commutative diagram for the causal abstraction. The divergence $d_{\tau,\omega}$ certifies consistency across the policy space, namely how close the predictions are when intervening-abstracting and abstracting-intervening.

Following Dyer et al. (2024), the *abstraction error* measures how closely the surrogate $\mathcal{M}'$ tracks the simulator $\mathcal{M}$ across interventions:

$$d_{\tau,\omega}(\mathcal{M}, \mathcal{M}') = \mathbb{E}_{\iota\sim\eta}\Big[d\Big(\tau_{\#}(\mathbb{P}_{\mathcal{M}_\iota}),\ \mathbb{P}_{\mathcal{M}'_{\omega(\iota)}}\Big)\Big] \tag{2}$$

where η is a distribution over the intervention set and d is a statistical divergence between the abstracted simulator distribution $\tau_{\#}(\mathbb{P}_{\mathcal{M}_\iota})$ and the surrogate's distribution $\mathbb{P}_{\mathcal{M}'_{\omega(\iota)}}$. The surrogate is α-*approximate* when $d_{\tau,\omega} \leq \alpha$; at $\alpha = 0$ the diagram in Figure 1 commutes exactly. When d is the Kullback–Leibler divergence, minimizing Eq. 2 reduces to maximizing the surrogate's log-likelihood on interventional data from the simulator (Dyer et al., 2024). We use the Variational ELBO as a tractable proxy for this objective.
The surrogate $\mathcal{M}'$ factors into *local* effects (infrastructure damage) and *network* effects (traffic) under the following assumptions.

Local Infrastructure Mechanism. Infrastructure damage in zone z depends only on the local state and actions:

$$I_{z,t} := f_{\text{infra}}(\text{Rain}_t, S_{z,t}, A'_{z,t}, \epsilon_I) \tag{3}$$

implying $I_z \perp\!\!\!\perp I_j \mid S_z, A'_z$ for $z \neq j$.

Global Traffic Mechanism. A flooded road in one zone causes delays in distant zones. Traffic outcomes depend on the *global* flood state:

$$L_{z,t} := f_{\text{delay},z}(\text{Rain}_t, \mathbf{S}_t, \epsilon_L) \tag{4}$$
$$C_{z,t} := f_{\text{cancel},z}(\text{Rain}_t, \mathbf{S}_t, \epsilon_C) \tag{5}$$

Traffic is conditionally independent of interventions (i.e. actions) *given* $\mathbf{S}_t$, reflecting the causal ordering: interventions → water depth → traffic.

4 System Architecture

To implement the coarse SCM, we employ Batched Multi-Task Gaussian Processes (GPs). GPs provide point estimates *and* epistemic uncertainty, crucial for robust planning under novel interventions. The architecture decomposes into three heads, each predicting one τ_Y while respecting the causal topology (Figure 2), namely the structure of the causal model.

Infrastructure Head For zone z and mode m: $I_{z,m} \sim \mathcal{GP}_{\text{infra},z,m}([\text{Rain}, S_z, \phi(A'_z)] \in \mathbb{R}^9)$, where $\phi(A'_z) \in \mathbb{R}^{K-1}$ encodes each categorical intervention as a vector of $K-1=7$ depth modifiers reflecting its physical effect on flood severity.

Travel and Cancellation Heads These ingest the full depth vector:
$L_{z,m} \sim \mathcal{GP}_{\text{delay},z,m}([\text{Rain}, \mathbf{S}] \in \mathbb{R}^{Z+1})$, and $C_{z,m} \sim \mathcal{GP}_{\text{cancel},z,m}([\text{Rain}, \mathbf{S}] \in \mathbb{R}^{Z+1})$. Network effects are captured through the input: each GP receives the city-wide flood state, so flooding in zone j can influence predicted delays in zone z.

We use sparse variational inference with P inducing points per GP, a Matérn-3/2 kernel with ARD, a Cholesky variational posterior, and a Softplus-warped Gaussian likelihood for non-negative outputs. Training maximizes the ELBO via mini-batch Adam Optimizer.

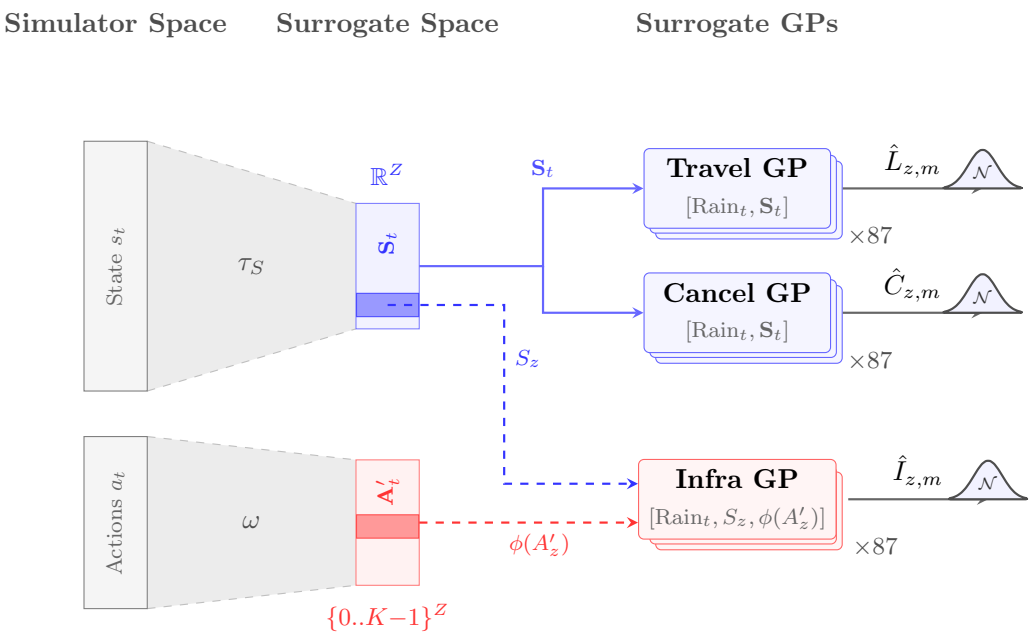


Figure 2: ICS Architecture. The MAAT state and actions are mapped through τ_S and ω to zone-level features. The Travel and Cancel heads receive the **global** flood vector $\mathbf{S}$; the Infrastructure head receives only **local** features (S_z, A'_z). Each head comprises $Z \times M$ sparse variational GPs producing Gaussian predictive distributions.

5 Case Study and Results

We focus on the inner city of Copenhagen, comprising $Z = 29$ TAZs and $M = 3$ transport modes (car, bicycle, on foot), giving 87 prediction tasks per head (261 GPs total). Each zone admits $K = 8$ categorical interventions (bioretention planters, soakaways, storage tanks, four permeable pavement variants, and a null action). The inner city was chosen for its location: particularly prone to flooding events and to the resulting disruptions.

To ensure broad interventional coverage, we collected training data from MAAT by executing a systematic grid of interventions across rainfall scenarios and action combinations, including edge cases of extreme rainfall and maximal deployment. This yielded $N_{\text{train}} = 308{,}000$ samples, with $N_{\text{val}} = N_{\text{test}} = 38{,}500$ each.

Each GP head is trained by maximizing the Variational ELBO via mini-batch Adam (batch size 256, learning rate 10^{-2}). We use $P = 300$ inducing points for the Travel and Cancel heads and $P = 150$ for Infrastructure. Inputs are standardized per-feature; outputs are min-max scaled per task. Total training time: ≈ 7.7 hours on a single GPU.

The ICS achieves high fidelity across interventional regimes. Table 1 summarizes test-set performance; Figure 3 shows the predicted-vs-true scatter plots.

Table 1: Test-set R^2 for the three ICS heads (Z=29, M=3, $N_{\text{test}} = 38{,}500$).

Head	**Input Type**	**Input Dim**	$\mathbf{R^2}$
Travel (Delay)	Global (Rain, $\mathbf{S}$)	30	0.978
Cancel (No-Travel)	Global (Rain, $\mathbf{S}$)	30	0.929
Infrastructure (Damage)	Local (Rain, $S_z, \phi(A'_z)$)	9	0.9999

The Infrastructure head achieves near-perfect accuracy ($R^2 = 0.9999$), expected for a structurally local process with low-dimensional inputs. The Travel head ($R^2 = 0.978$) shows that the global-input architecture captures network-level coupling. The Cancel head ($R^2 = 0.929$) is hardest—trip cancellations are sparse events—yet the model achieves

strong accuracy.

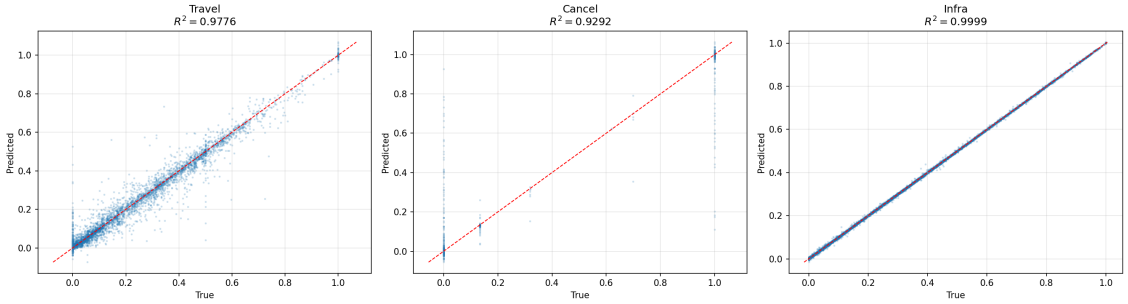


Figure 3: Predicted vs. true values on the held-out interventional test set. Strong alignment along the identity line indicates low abstraction error $d_{\tau,\omega}$.

6 Discussion and Future Work

Our results sugest that an ICS can replace simulators for policy evaluation. A black-box machine learning model that instead ignores the underlying causal structure would instead see its performance degrade as soon as a policy intervention causes domain or out-of-distribution distributional shifts. In the specific case of the inner city, the embedded causal structure explicitly considers the distinction between local and global effects which, in turn, allows to decouple the estimation of infrastructure damage from trip cancellation and travel delays. Besides, the GP posterior distribution allows to quantify the uncertainty associated to each surrogate run, which in turn would allow to identify which interventions cause such shifts that even an ICS may fail to capture. This is very important in machine-learning supporting policy-making as the level of trustworthiness of the surrogate can be quantified, allowing the planner to reject interventions that are related to unacceptable levels of uncertainty.

Still, there are some limitations that may be tackled in future research directions. First, as mentioned in Section 5, the surrogate shows slightly lesser performance in estimating trip cancellations as these events are not continuous but instead characterized by a boolean threshold (the trip is either canceled or is not). This suggests that the GP model may struggle with similar outputs but more experiments are needed to confirm the assumption, as the R^2 is still quite high, even for cancellations (0.929). Besides, different surrogates should be tested in the proposed ICS framework, to ensure the widest applicability of the approach. A natural next step is integrating the ICS into a Reinforcement Learning (RL) loop, where the surrogate replaces the simulator during policy training — addressing the fundamental incompatibility between the simulator's cost and RL's demand for millions of evaluations. Preliminary results on policy consistency between surrogate-trained and simulator-trained agents are encouraging and will be presented in a follow-up study. Scaling to the full Copenhagen network (∼280 zones) is also a priority.

7 Conclusion

By formalizing surrogate modelling through causal abstraction (Dyer et al., 2024), we have developed a framework that is both computationally efficient and interventionally consistent. The decomposition into local and global GP heads—dictated by the causal structure of the physical system—achieves $R^2 > 0.92$ across all heads on a held-out interventional test set. The ICS framework provides a principled foundation for accelerating iterative planning in urban flood resilience, with extensions to policy optimization and RL.

Acknowledgements

The work was supported by Novo Nordisk Foundation grant NNF23OC0085356.

References

Anand, T. V., Ribeiro, A. H., Tian, J., & Bareinboim, E. (2023). Causal effect identification in cluster DAGs. In *Aaai conference on artificial intelligence*.

Beckers, S., & Halpern, J. Y. (2019). Abstracting causal models. In *The thirty-third AAAI conference on artificial intelligence* (p. 2678-2685). (arXiv:1812.03789 [cs.AI]) doi: 10.1609/aaai.v33i01.33012678

Christiansen, H., & Baescu, O. (2024). *The danish national travel survey: 0623v1.* `https://doi.org/10.11581/dtu:00000034`. (Accessed: 2024-06-01)

Costa, M., Vandervoort, A., Schmidt, C., Miranda, J., Petersen, M. W., Drews, M., ... Pereira, F. C. (2026). *Artificial intelligence for climate adaptation: Reinforcement learning for climate change-resilient transport.* Retrieved from `https://arxiv.org/abs/2603.06278`

Danmarks Meteorologiske Institut. (2023). *Klimaatlas.* Retrieved from `https://www.dmi.dk/klima-atlas/data-i-klimaatlas` (Accessed: 2024-08-26)

Dyer, J., Felekis, Y., Bishop, N., Zennaro, F. M., Damoulas, T., Calinescu, A., & Wooldridge, M. (2024). Interventionally consistent surrogates for complex simulation models. *38th Conference on Neural Information Processing Systems (NeurIPS 2024).*

Ge, Q., Ciuffo, B., & Menendez, M. (2014). An exploratory study of two efficient approaches for the sensitivity analysis of computationally expensive traffic simulation models. *IEEE Transactions on Intelligent Transportation Systems*, *15*(3), 1288–1297.

Kekić, A., Schneider, J., Büchler, D., Schölkopf, B., & Besserve, M. (2025). *Learning nonlinear causal reductions to explain reinforcement learning policies.* arXiv preprint arXiv:2507.14901.

Madaleno, F., Misra, P., & Markham, A. (2026). *Coarsening causal dag models.* Retrieved from `https://arxiv.org/abs/2601.10531`

Natterer, E., Engelhardt, R., Hörl, S., & Bogenberger, K. (2025). *Machine learning surrogates for optimizing transportation policies with agent-based models.* (https://arxiv.org/abs/2501.11057)

Osorio, C., & Bierlaire, M. (2013). A simulation-based optimization framework for urban transportation problems. *Operations Research*, *61*(6), 1333-1345.

Pearl, J. (2009). *Causality* (2nd ed.). Cambridge University Press.

Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference: Foundations and learning algorithms.* The MIT Press. Retrieved from `https://web.math.ku.dk/~peters/jonas_files/ElementsOfCausalInference.pdf`

Pregnolato, M., Ford, A., Wilkinson, S. M., & Dawson, R. J. (2017). The impact of flooding on road transport: A depth-disruption function. *Transportation Research Part D: Transport and Environment*, *55*, 67-81.

Rubenstein, P. K., Weichwald, S., Bongers, S., Mooij, J. M., Janzing, D., Grosse-Wentrup, M., & Schölkopf, B. (2017). Causal consistency of structural equation models. In *Conference on uncertainty in artificial intelligence (uai).*

SCALGO. (Accessed:2024-06-07). (https: //scalgo.com/live/denmark)

Tang, C. J., Z., Zhou, Z., Zhou, T., & Yuan, K. (2025). The sensitivity analysis of parameters in the 1d–2d coupled model for urban flooding. *Applied Sciences*, *15*(4), 2157.

Vejdirektoratet. (2022). *Grøn mobilitetsmodel (gmm).* `https://www.vejdirektoratet.dk/segment/groen-mobilitetsmodel`. (Accessed: 2024-06-06)