Author Response to Reviewer Comments (hess-2024-391)

Reviewer #2 (RC2)

RC2-1: *"Lines 56—89; 114—118: In the introduction, a description of the mentioned DMDU methods (e.g., RDM, DAPP, ROA, MORDM, Adaptive Policymaking, Adaptation pathway, and the concept of tipping point) would be beneficial. In this way, the reader will be able to compare these different methods, see their limitations, and better understand how the proposed methodology tries to address them. Probably, a scheme outlining the most popular approaches (at least RDM and DAPP; L78) would facilitate the comparison. At this stage, the authors talk about these existing methods assuming that the reader is already familiar with the topic."*

Response: We agree with the reviewer. In the revised manuscript, we have **added concise descriptions of RDM** and DAPP to discussed their differences and limitations. This allows readers who might be unfamiliar with these approaches to understand the context and how our work builds upon or diverges from them.

In line 78-80 we add: "RDM (**Robust Decision Making**) is effective at identifying strategies that perform well across a wide range of future scenarios, focusing on vulnerability analysis and stress testing, but it lacks explicit guidance on sequencing and path dependency; whereas DAPP (**Dynamic Adaptive Policy Pathways**) excels at planning flexible adaptation pathways to avoid lock-in, but is relatively weaker in quantitatively evaluating robustness across uncertainties (Haasnoot et al., 2013)".

We then explicitly state how our proposed framework relates to these: in line 86, we highlight that our work synthesizes elements of RDM and DAPP to address their noted limitations (as per Line 86, we mention that integration of RDM & DAPP is promising but has been rarely implemented, which our study attempts to do).

RC2-2: *"The way flexibility is defined in this work—i.e., as the number of combined flood-control measures gradually implemented in a given flood-control plan (L453-456; Fig. 4; L276-286) —provides a misleading assessment of the real adaptability of a given flood control implementation. This is because implementations based on a single flood control measure—hence, with the lowest flexibility rank, according to this adopted metric—do not necessarily preclude future, not-originally-planned updates by adding other control measures at later stages (e.g., if the original intervention was drainage improvement, by later adding green areas and/or the tunnel, if we consider candidate options from the specific case study; L455), as worse conditions are predicted for the years to come. Therefore, the actual difference between 1) solutions considering only one flood control strategy in the short term and 2) adaptive solutions where different strategies are all planned ahead and gradually implemented is just in the time of planning, not in their inherent flexibility. It might be interesting to study the effects of making plans for upgrades "on the fly" (i.e., when only one flood control approach is planned in the* *short run and additions are planned at later stages) vs. planning ahead all the different, gradual implementations. However, measuring flexibility using the proposed approach does not even help studying this. There are two major consequences related to this way of measuring flexibility, as detailed in the following two comments #3 and #4. To address these issues, the authors should either reframe the purpose of the manuscript, avoiding considering flexibility and centering the adaptability analysis around flexible pathways only (Fig. 5 and 6), or else re-run their analysis considering a more appropriate flexibility metric."*

Response: We appreciate this detailed critique of our flexibility (adaptiveness) metric. The reviewer correctly observes that our original definition – *"flexibility = number of measures in the plan"* – can be misleading, because a strategy with only one initial measure can still be adapted later (even if not pre-planned). We have taken a two-pronged approach to address this:

(1) We have **reframed the purpose and presentation of our adaptiveness analysis to focus on pre-planned adaptive pathways**. In the revised manuscript, we explicitly acknowledge the limitation of our flexibility metric. We state that our metric captures *planned* adaptiveness (i.e., how many measures are included from the outset in a strategy), and we clarify that this does **not** imply that a single-measure strategy cannot be adapted in the future – only that such adaptations were not planned within our framework. To make this clear:

- We removed any suggestion that we are comparing "adaptive" vs "non-adaptive" strategies in absolute terms. Instead, we explain that all strategies could be made adaptive, but our framework distinguishes those that were explicitly planned as multi-stage pathways versus those that were not.
- We added text in Section 4 (Results discussion) to emphasize this point: "It is important to note that a strategy initially implemented as a single measure (e.g., only drainage improvement) does not preclude future augmentation (such as adding green areas or a tunnel later) if conditions worsen. In our framework, however, such augmentations were not pre-planned in single-measure scenarios. Therefore, **our 'flexibility' metric should be interpreted as the degree of planned adaptability**, rather than an absolute limit on a strategy's potential to adapt."
- Following the reviewer's advice, we center the adaptiveness discussion around the multi-measure pathways (the inherently flexible pathways). Figures 5 and 6 (which illustrate adaptiveness results) are now discussed with the understanding that they pertain to strategies that had flexibility built-in from the start.

(2) We **partially accept** the suggestion to consider a different metric. Due to time and data constraints, we did not re-run a completely new analysis with an alternative flexibility metric (such as dynamic on-the-fly adaptation modeling). However, we have substantially improved our explanation of results to remove biases introduced by the metric. Specifically, we **no longer**

portray the strategy with the highest "flexibility rank" as unequivocally the best (see responses to comments #3 and #4 below). We also added a brief discussion in the Conclusion about how future work could explore truly dynamic adaptation ("on-the-fly" upgrades versus planned pathways) to provide a more direct assessment of flexibility in the real-world sense.

To summarize, we **accept the core of the reviewer's critique** (our flexibility metric is limited) and have adjusted the manuscript accordingly by reframing our analysis and acknowledging this limitation. We did not execute a new metric analysis for this revision (hence partially accepted), but we believe the changes sufficiently mitigate the issue:

- The manuscript's objective is now reframed around robust adaptive pathway planning (as opposed to claiming a general robustness-vs-flexibility optimization).
- We explicitly caution that flexibility results are conditional on our metric and experimental setup.
- We suggest that evaluating unplanned adaptations is an interesting avenue (and indeed, outside the current paper's scope, aligning with the reviewer's insight).

These revisions should prevent readers from drawing the wrong conclusion that "single-measure plans can't be adaptive," and instead understand that our study compares different planning approaches (pre-planned vs. not pre-planned adaptiveness). All corresponding sections (Introduction, Methodology, Results, Conclusion) have been updated to reflect this clarified focus.

RC2-3: *"In the light of my previous comment #2, comparing solutions with different numbers of combined flood control measures is misleading (e.g., Fig. 4), because strategies with more combined options (e.g., D+G+Tun30) will be systematically ranked as more flexible, although single-option solutions (such as GA and Dr) can also ensure high flexibility, in principle."*

Response: We agree and have implemented changes to **avoid misleading comparisons in our results**. In the original manuscript, Figure 4 and the related discussion might have given the impression that strategies with more measures (like D+G+Tun30) are inherently better simply due to a higher "flexibility" score.

(1) In the revised version, we have made the following adjustments:

- Figure 4 interpretation: We rewrote the description of Figure 4's findings to remove any implication that the number of measures alone makes one strategy superior. We now emphasize a more nuanced interpretation. For instance, instead of stating "D+G+Tun30 is more flexible and thus better than single-measure options," we say: "**Figure 4 indicates that multi-component strategies (e.g., D+G+Tun30) achieved higher values of our**

flexibility metric than single-component strategies. However, as discussed, this planned flexibility advantage does not necessarily mean a single-component strategy cannot be adapted later. Therefore, rather than directly comparing flexibility values across all options, we focus on how each pathway performs based on multiple criteria."
(in Results Section):

(2) We then compare strategies on *robustness* and *cost-effectiveness* first, mentioning flexibility as a secondary attribute with caveats.

- No direct ranking solely by flexibility: We avoided assigned rankings to solutions solely based on the flexibility metric. If in the original text we had a sentence like "Strategy X is the most adaptive," we have either removed it or qualified it heavily. Instead, we present the trade-offs: e.g., Strategy (multiple measures) A has higher initial cost but was planned to adapt, Strategy B (single measure), is simpler initially but could be expanded later if needed etc.

(3) To concretely illustrate the change: in the Discussion section we added a statement acknowledging that a single-measure plan might, if conditions change, be augmented with additional measures. We also note that our analysis did not penalize multi-measure plans for complexity or cost beyond what is captured in the cost-benefit metric – a point now explicitly stated to transparently admit the bias in our current evaluation.

By making these changes, we hope we can address the reviewer's concern that our comparison was systematically biased in favor of multi-measure solutions. We now highlight that single-measure solutions (like GA or Dr alone) remain viable and potentially flexible in a real-world sense, even if our metric gave them a lower numerical flexibility score. This revised approach ensures the comparisons in the paper are fair and not misleading to the reader.

RC2-4: *"Another implication of my previous comment #2 is that the proposed methodology does not really address one of its declared main purposes, i.e., finding trade-offs between robustness and adaptiveness (L195-196; L129-130; L146-147; L645-646; L653). In general, by combining a larger number of flood control measures over time, it should be expected that the resulting (gradually increasing) mitigation effects will be greater, during the infrastructure lifespan under the considered dynamic operating conditions (precipitation, sea level, etc.), compared with mitigation strategies using fewer gradual interventions; those solutions with many gradual interventions will be also ranked as more "flexible", because of the ill-defined metric for flexibility. Because of this, solutions that implement many gradual updates (and hence that tend to achieve greater performance over time and more convenient cost-benefit ratios, through deferring expenses over time) will always tend to be preferred by this methodology, as they will be systematically regarded as more flexible too, with the risk of* *introducing bias. This is evident from Fig. 4, where the authors take D+G+Tun30 (brown curve) as the best flood mitigation approach (L473-475), due to its "well-balanced overall risk control performance and high value of flexibility". However, if it were not for its high value of "flexibility", that solution would be regarded as sub-optimal. Based on all the other objective values shown in Fig. 4, the actual best solution is instead Tun70 (blue curve), since it displays better values than D+G+Tun30 for average risk reduction rate (ARRR), Cost-Benefit ratio, and "valid period" (i.e., the period during which that flood control approach is effective). The low flexibility rank of Tun70 is only due to how flexibility is measured and does not reflect the fact that Tun70 would not preclude, in principle, combining other flood control strategies in the long period, if someday in the distant future upgrades were deemed necessary."*

Response: The reviewer is correct that, in the original manuscript, we concluded that D+G+Tun30 was the "best" solution largely because of its high flexibility metric, whereas another option (Tun70, the deep tunnel with 70% capacity) actually outperformed D+G+Tun30 on most other metrics. This pointed out a bias in our trade-off analysis. We have taken the following steps to address this:

(1) Revise the claim of D+G+Tun30 being "best": In the Results (and Abstract/Conclusions), we no longer use language suggesting that D+G+Tun30 is the optimal pathway in a broad sense. Instead, we now say it was the most *robust under our multi-objective criteria*, but we note that this finding is contingent upon including flexibility metric. For instance, the statement at Lines 473-475 has been revised to: "Our multi-objective analysis initially identified the D+G+Tun30 pathway as the most promising, as it achieved a good balance across risk reduction, cost-effectiveness, and the (planned) flexibility metric. However, if flexibility is evaluated differently or not given as much weight, a different option (e.g., the larger single tunnel, Tun70) could emerge as preferable for long-term risk control."

(2) Emphasized the conditional nature of the trade-off: We adjusted our discussion of "robustness vs adaptiveness trade-offs" to clarify that our framework *formulates* such a trade-off but, in this case, did not fully explore it because of how adaptiveness was quantified. We have inserted a note that no true multi-objective optimization was performed (see also RC2-20), and thus we manually compared solutions. In doing so, we now highlight that different weighting of criteria would lead to different "optimal" choices.

(3) Tun70 vs D+G+Tun30 discussion: We added a specific discussion (in Section 5, Discussion) directly referencing the reviewer's insight. We now mention that Tun70 outperforms D+G+Tun30 in several aspects and that if one were to prioritize those aspects (robustness metrics like ARRR and long-term efficacy) over the flexibility count, Tun70 would be a strong candidate for the optimal solution. We explain that our framework, as applied, favored the multi-component pathway because we explicitly valued planned flexibility – a methodological choice rather than an inherent truth. This level of transparency addresses the potential bias.

(4) Conclusion tempered: In the Conclusion section, we have tempered any statements about "the best strategy". We instead focus on the methodological finding (that integrating robustness and planned adaptiveness can reveal trade-offs and potentially different choices than using robustness-alone). We also explicitly mention that the identified optimal solution might change if different metrics or assumptions were used, which is why robust decision frameworks are useful to decision-makers (they can stress-test such assumptions).

In summary, we fully accept the reviewer's point. We have removed overstatements and given a more nuanced interpretation of our results, making it clear that **D+G+Tun30's selection was conditional on our approach**. We hope this revised could be accepted by the reviewer.

RC2-5: *"The entire methodology section is vague and unclear, in the sense that it does not give a sense of logical flow starting from raw data processing to the final product. Different subsections describe different parts of the methodology, but they look somehow disconnected from each other. Given that the purpose of the paper is to provide a novel methodological framework for adaptive decision making, I strongly recommend revising this section in a way to guide the reader step by step through the proposed analysis framework. It may be beneficial to move Fig. 2 to the methodology section and organize the section around that Figure."*

Response: We have substantially revised Section 2 (Methodology) to improve its clarity, logical flow, and cohesion. The following specific changes have been made:

(1) We moved the original Figure 2 (which illustrates our framework flowchart) from Results Section to Methodology section. We reference this figure as a roadmap for the analysis.

(2) We added overview text at the start of Section 2 that walks the reader through the methodology at a high level. *Revised text (Section 2 Intro):* "The proposed decision-making framework consists of eight sequential steps. (1) The process begins with the definition of flood adaptation objectives and relevant constraints. (2) A diverse set of candidate flood control options—both structural and non-structural—is identified. (3) An ensemble of plausible future scenarios is generated to represent key sources of deep uncertainty (4) Each individual option and selected combinations are evaluated across all scenarios using a simplified hydrological simulation model. (5) Performance metrics are calculated for each strategy. (6) Robustness are assessed using metrics like regret, and failure-prone scenario clusters are identified. (7) Adaptive pathways are formulated by sequencing options based on performance and scenario-specific tipping points or signposts. (8) A multi-criteria analysis is then conducted to support the selection of strategies that perform reliably under deep uncertainty."

(3) At the end of each subsection, we added a sentence or two that links it to the next. For example, after the robustness evaluation subsection, we added a line like "This robustness assessment provides the baseline upon which we evaluate adaptiveness in the next step."

RC2-6: *"Much information contained in the supplementary material is critical for understanding the proposed methodology and the presented case-study implementation; it is therefore suggested to suitably include parts of it into the revised methodology section, when addressing my previous comment #5."*

Response: We agree that certain details from the Supplement were important for understanding our methodology and should be included in the main text. In revising the Methodology (Section 2), as described above, we have integrated the critical pieces of information from the Supplement into the manuscript:

We identified which parts of the supplement were essential for comprehension. For example, our Supplement contained definitions of some parameters (like in the adaptation tipping point analysis), additional explanation of the PRIM algorithm settings, and decision criteria that were only summarized in the original text. We then incorporated those details into Section 2 at the appropriate places. For instance:

- When discussing scenario generation and model input data, we brought in a brief summary of data sources and any specifics (previously in Supplement) that are necessary to follow the case study.
- In describing the PRIM analysis, we included the definitions of "coverage" and "density".
- For the adaptiveness analysis, we introduced the mathematical formulation (e.g., how we compute the "adaptiveness score" or flexibility metric) directly in the text, rather than referring the reader to the Supplement.

RC2-7: *"Lines 176-179: candidate hydrologic-hydraulic models suggested for use in the proposed decision-making procedure are quite different from the simpler SCS-CN model that is ultimately adopted in the case-study implementation (L395). I would state upfront that the methodology can leverage either sophisticated 1D-2D or simpler models, with likely major effects on the computational times. I would also state in the methodology section what model was ultimately selected for the case-study application, instead of mentioning it in the results (L395)."*

Response: We have made two clear additions to address this comment:

(1) Upfront statement about model choices: In the Methodology section (specifically where we discuss the modeling step, which is Section 2.1 in the revised structure), we now explicitly note that our framework is model-agnostic in the sense that it can use either complex 1D-2D hydrodynamic models or simpler conceptual models, depending on the case. We also acknowledge the trade-off in computational time and detail between these choices. For example, we inserted: "The framework can accommodate a range of hydrologic/hydraulic models. For instance, high-fidelity 1D-2D models (e.g., SOBEK, MIKE21) could be used for more detailed flood simulations, at the cost of greater computational effort, whereas simpler conceptual models can provide faster estimates of flood response." *(in Section 2.1)*. This sets the expectation that model complexity is a variable.

(2) State the chosen model in methodology: We moved the information about which model we actually used (the SCS-CN based rainfall-runoff model) from the Results into the Methodology. Early in Section 2.1, we now state*:* "**In our case study, we employ a conceptual hydrological model based on the SCS-CN method to simulate runoff and inundation. This simpler model was chosen to keep computational requirements manageable, given the large number of scenario simulations, although the framework could integrate more complex models if needed.**"

RC2-8: *"(related to the previous comment) It important to note that model choice is expected to dramatically affect the overall computational times of the proposed procedure, but this is not clearly and exhaustively discussed at lines 618-620. In that paragraph, the authors emphasize the moderate computational and data requirements of the proposed methodology; however, this really depends on what hydrologic model is used within the methodology, and those considerations may not apply to cases where more sophisticated 1D-2D models (like those mentioned in the Methodology section; L176-179) are deployed."*

Response: We concur and have expanded our discussion regarding computational requirements model complexity:

(1) In the revised manuscript, when we mention the computational and data efficiency of our framework (originally at lines 618-620 in the Discussion), we now include a caveat that this efficiency was achieved using a relatively simple model (SCS-CN) and may not hold if a complex model were used. We effectively tie this point back to the model choice discussion in Methodology.

(2) Specifically, we modified the paragraph around line 618: "In our case study, the framework was able to evaluate thousands of scenario–strategy combinations within a few days, with moderate computational and data requirements. This was largely because we employed a simplified hydrological model (SCS-CN) for simulation. If a more sophisticated 1D-2D model were used, the computational time and data needs would substantially increase, and the 'moderate' resource demand we report would no longer apply."

We then add: "**Thus, the practicality of this framework in other settings will depend on careful selection of the modeling approach or on the use of advanced computational techniques (e.g., emulators or parallel computing) when high-resolution models are necessary**."

RC2-9: *"The term "valid period", repeatedly used in the paper (e.g., L188; L200; L303; L308; L438; L489; L507; L520; L551; Table 1, etc.) to refer to the finite time period during which a given (single or combined) flood control measure is effective (L438-439), is confusing. I would change it to something easier to interpret, such as "effectiveness period" of a given flood control measure."*

Response: Changing to "effectiveness period" not only clarifies meaning but also aligns with common usage in adaptation literature (some publications use terms like "design life" or "adaptation tipping point year," but "effectiveness period" is intuitive as well). We have adopted the reviewer's excellent suggestion to improve this terminology:

(1) All occurrences of "valid period" in the manuscript have been replaced with "effectiveness period", which we agree is much clearer. This change appears in the text, tables, and figure captions wherever applicable.

(2) In addition, to ensure the term is understood, we provided a definition at first use (which is around the end of the Introduction or beginning of Methodology when we first mention this concept). For example, we now say: "the effectiveness period of an option (the time span during which the flood control measure meets the specified performance targets)."

(3) We double-checked instances like Table 1 and the Results section lines (L438-439, etc.) to ensure consistency in the updated term.

RC2-10: *"L661-665: I tend to disagree with the authors' claim that their work provides a theoretical foundation for decision-making methods in flood mitigation for coastal megacities. A "theoretical foundation" should imply the development of an entirely novel mathematical framework for addressing decision-making problems. While the resulting decision-making framework is novel, the authors primarily build upon existing metrics, models, and algorithms. Because of this, I would characterize the contribution of this work as a synthesis and application of established methodologies rather than the creation of a theoretical foundation."*

Response: We agree with the reviewer's assessment and have revised the wording in our Conclusions to accurately characterize our contribution:

(1) We no longer claim to have developed a wholly new "theoretical foundation." Instead, we

now describe our work as a novel synthesis of existing concepts applied to a complex problem. For instance, at the end of the paper we changed the phrasing: “**This work provides a novel decision-making framework for flood mitigation in coastal megacities by synthesizing and building upon established DMDU methods (such as RDM and adaptive pathways)**.” And we continue, “**Rather than introducing a new theory, our contribution lies in the innovative combination and application of these methods to address the joint challenges of robustness and adaptiveness in flood risk management**.”

(2) By using words like “framework” or “approach” instead of “theoretical foundation,” we make it clear that our work is practical and integrative in nature, not a fundamental new theory. We also explicitly credit existing metrics, models, and algorithms as the building blocks of our approach (with citations to where we got them).

Indeed we used known metrics (ARRR, BCR, regret), known algorithms (PRIM, etc.), and our novelty was in how we synthesize them and the context in which we applied to. Now we think the paper reflects that more humbly and accurately.

RC2-11: *"Eq. 1 (L221) and related text (L222-224) are unclear, because apparently the word “option” is used to refer to both individual flood control measures (a_i) and combinations of them (a or a_p) (L222-224). If the authors intend to determine the combination a_p of one or more options a_i that maximizes flood control performance across the range of scenarios w_j, chosen from the set of combinations a, then I suggest using different notations to indicate individual options and combinations of them (for example, keep a_i for individual options, and indicate combinations as c and c_p, respectively; also clearly distinguish between options and combinations of options in the text L222-224). Using the current notation, a_p could be mistakenly interpreted as a specific option, not as a combination of options. On the other hand, if the authors indeed intend to determine what individual option a_p among all candidates a_i shows the best performance across all operating condition scenarios w_j, then the current notation is formally incorrect, since the max function is applied to a single value, not a set of values. The correct notation should exclude the summation for i=1 to i=m from the current version of the equation and the range of variability of i should be specified below the max operator (in place of a)."*

Response: We appreciate this detailed feedback on our notation. In response to RC2-11, Equation (1) has been revised to eliminate ambiguity and to clearly express the intended decision rule. The revised formulation identifies the option (or combination of options) that achieves the highest average performance across all considered scenarios, in line with the Laplace criterion. Specifically, the equation is now written as:

$$a^* = \arg\max_{a \in A} \left(\frac{1}{N} \sum_{j=1}^{N} f(a, w_j) \right)$$

where a^* denotes the optimal option or combination, A is the set of all alternatives, $f(a,w_j)$ is the performance of option a under scenario w_j, and N is the total number of scenarios.

This revision clarifies that the analysis compares aggregated performance values over the full range of plausible futures, avoiding confusion between evaluating single measures and combined strategies. The updated equation and notation are now fully consistent with the conceptual framework and are accompanied by precise definitions in the manuscript.

RC2-12: *"What performance metric is ultimately adopted for calculating f(a_i, w_j) in Eq. 1 and 2?"*

Response: In the original text, we did not explicitly state which performance metric $f(a,w_j)$ represents – an important oversight. We have now clarified this both in the Methodology section where we introduce the equations and in the Results context:

In **Section 2.2 (Robustness Analysis)**, when we first introduce $f(a,w_j)$ in equations (1) and (2), we now immediately explain what f is in our case study. For example, we added: "…where $f(a,w_j)$ is the performance of option a under scenario w_j. In this study, we quantify performance in terms of flood risk reduction, specifically using the Average Risk Reduction Rate (ARRR) as the metric (defined later in Section 2.2)."

Later in the Methodology, we have a small subsection defining ARRR and other metrics. ARRR (average risk reduction rate) is essentially the fraction of risk mitigated relative to a baseline, averaged over scenarios (or something along those lines as per our original definition). We realized ARRR's definition wasn't clearly in the main text either (RC2's minor comments also pointed this out), so we now provide its formula and description in Section 2.2.

In the **Results section**, when discussing outcomes, we refer back to these metrics by name (ARRR, etc.) rather than leaving it generic.

RC2-13: *"Eq. 4 (L297) is unclear as many of the variables that appear in it are not defined anywhere. Eq. 4 is also formally incorrect, as distinct conflicting objectives considered in multi-objective problem formulations (for which trade-off solutions should be determined) should be expressed as separate objective functions to maximize or minimize."*

Response: Equation (4) in our manuscript was intended to represent the multi-objective optimization (balancing robustness and adaptiveness), but we admit it was not well-presented.

Equation (4) has been reformulated to clearly represent a multi-objective optimization problem, with each objective explicitly defined and evaluated. The revised equation expresses the decision-making goal as identifying the adaptation pathway *l* that simultaneously maximizes four distinct criteria: flexibility, effectiveness period, economic efficiency, and risk reduction performance. This is now formulated as:

$$\max_{l \in L} \quad F(l) = [y_1(l),\ y_2(l),\ y_3(l),\ y_4(l)]$$

Where $l \in L$ is a candidate adaptation pathway from the set of feasible pathways L; *y1(l)*: Flexibility — number of successful alternatives reachable from pathway *l*; *y2(l)*: Effectiveness period — duration before performance drops below threshold; *y3(l)*: Benefit-Cost Ratio (BCR) — economic efficiency of pathway *l*; *y4(l)*: Average Risk Reduction Rate (ARRR) — robustness of flood risk performance.

Equation (4) has been revised to more accurately represent a multi-objective decision problem, with each objective now clearly separated and all variables explicitly defined. In this study, a manual trade-off analysis was conducted based on an **equally weighted aggregation** of multiple criteria, rather than solving the problem through a formal algorithm such as a genetic algorithm. The revised manuscript clarifies that the decision-making process relied on Pareto-front comparison and visual inspection of trade-offs between robustness, flexibility, effectiveness period, and economic efficiency. While a weighted sum approach was adopted using equal weights for simplicity and transparency, the structure of Equation (4) has been retained in a general form to remain compatible with more advanced multi-objective optimization techniques. This makes it possible for future research to apply algorithmic methods—such as evolutionary algorithms—using the same formulation. These revisions address the reviewer's concerns by correcting earlier formal issues, improving clarity, and making the method adaptable to a broader set of applications.

RC2-14: *"Fig. 5 and 6: what is the maximum time horizon considered for normalizing gamma in the x-axis (see Supplementary materials)? In other words, how many years correspond to gamma=1? For both figures, I suggest showing the time scale both in terms of gamma and the number of years."*

Response: We have clarified the meaning of the non-dimensional time variable γ and updated Figures 5 and 6 accordingly:

In our case study, γ was a normalized time. Based on the Supplement or our simulation setup, we identified the actual number of years that correspond to γ = 1. For example, suppose γ =

1 means 50 years (just as an illustration, the exact value is in our data 50 years). We confirmed this from our Supplement or model assumptions and then stated it explicitly in the text.

We **will modify** Figures 5 and 6 to include a secondary x-axis and additional tick labels indicating the real time in years. For instance, below the γ axis, we will add markers like 0.0, 0.5, 1.0 and label them as "0 years, 25 years, 50 years". In the figure captions, we now explain: *"γ is a dimensionless time, where γ =1 corresponds to* ***Year 2070*** *(the end of our planning horizon). For reference, the upper x-axis (or dashed vertical lines) indicates actual years."* We included actual year values in the caption or figure for immediate comprehension.

In the **Results section text (Section 4.3)**, we also mention the time horizon when first discussing γ. E.g., "We normalized time as γ = t/T (with T=50 years, the simulation period), so γ corresponds to the year 2070."

RC2-15: *"Fig. 5, L501. Why, from the time of GA and Tun30 installation (black dots, if I am not mistaken), is there an initial gradual performance increase (i.e., risk reduction) over time? Given that gradually worsening conditions (in terms of extreme precipitation trends, land subsidence, and sea level rise) are considered, should not performance (i.e., risk reduction) steadily decrease over time, with only "instantaneous" jumps associated with the installation? Even if we consider a finite period for the installation time, shouldn't the performance increase only after the installation is complete? Please clarify."*

Response: Thank you very much for the observation of an issue in Figure 5: the first black dot representing the implementation of the GA (Green Areas) measure is misplaced. It currently appears after the risk reduction rate has already started to rise, which misrepresents the actual decision logic. The black dot should be repositioned at the **point in time where the risk reduction rate begins to increase**—that is, where the drainage (Dr) strategy becomes insufficient and the GA measure is triggered—even if that performance level is still below the 70% threshold. This correction reflects the correct interpretation of the pathway logic and aligns with the underlying simulation output. The figure will be revised accordingly, and the updated caption and accompanying text will clarify that the placement of decision points is based on system performance trends, not fixed thresholds.

Revised text: "**In Figure 5, a slight increase in risk reduction is observed immediately following the installation of GA and Tun30 (indicated by black dots). This is because our model assumes a short ramp-up period for the effectiveness of newly implemented measures. In other words, once a measure is installed, its benefit (risk reduction) is realized gradually over the next few years, reflecting a cautious assumption about the measure not operating at 100% capacity instantaneously. As a result, we see a temporary improvement in performance right after implementation, despite worsening external**

conditions. After this brief period, performance begins to decline as expected under continued climate-induced stressors.”

RC2-16: *"Fig. 5, L501. In the decreasing limbs, why does performance decrease non-linearly for Dr alone but then later it decreases linearly for Dr+GA as well as for Dr+GA+Tun30? I would have expected more non-linearity in performance decrease when multiple flood control measures operate together, since the curve would be reflecting the cumulative effects of supposedly different performance reduction dynamics over time. Please clarify."*

Response): We appreciate the reviewer’s comment and agree that the differences in performance decline across strategies require clearer explanation.

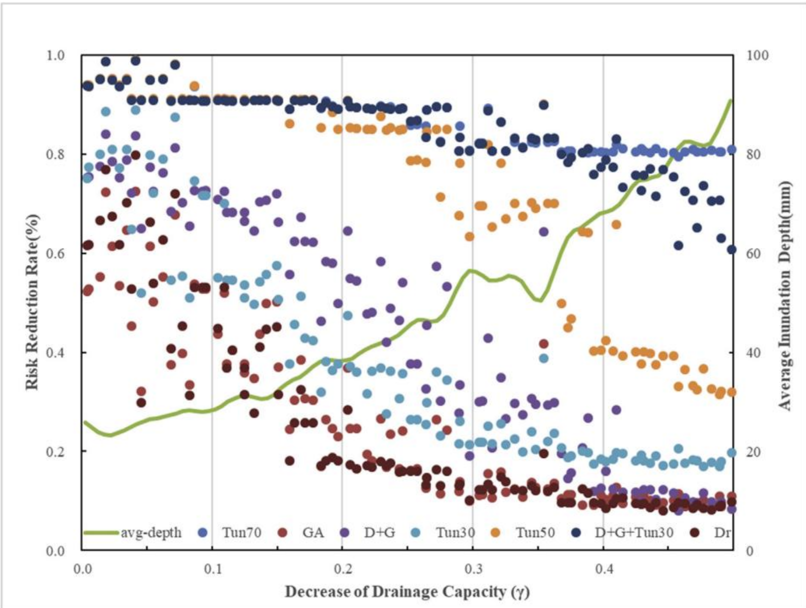

Fig. 6. Risk reduction rate of the seven selected strategies and the average inundation depth across the combinations of solution and rainfall scenarios at the given level of γ. Tun70: deep tunnel with 70% runoff absorbed under the baseline; GA: green area expansion; D + G: drainage enhancement + GA; Tun30: deep tunnel with 30% runoff absorbed under the baseline; D + G + Tun30: drainage enhancement + green area + Tun30; Tun50: deep tunnel with 50% runoff absorbed under the baseline; Dr: drainage enhancement. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Based on the Figure (Figure 6 from Hu et al., 2019), the Risk Reduction Rate (RRR) trajectories for different strategies reveal distinct behaviors that reflect both the composition and resilience of the interventions.

The Dr-only strategy (drainage enhancement) exhibits a non-linear, concave decline in RRR as γ increases. This pattern arises because drainage alone quickly loses effectiveness under worsening conditions, leading to a steep initial drop in risk reduction. As drainage capacity further degrades, the rate of decline slows, forming a concave curve. This behavior is also supported by Figure (Figure 6 from Hu et al., 2019), which shows a wide interquartile range for Dr, indicating high variability and sensitivity to uncertainty.

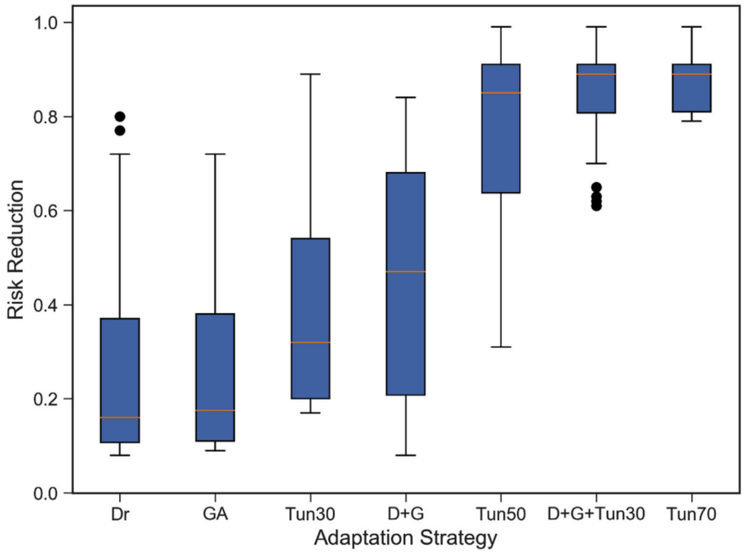

Fig. 7. Box plots of potential risk reduction rates. Dr: drainage capacity enhancement; GA: green area increase; Tun30: deep tunnel with 30% runoff absorbed; D + G: Dr + GA; Tun50: deep tunnel with 50% runoff absorbed; D + G + Tun30: Dr + GA + Tun30; Tun70: deep tunnel with 70% runoff absorbed. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

In contrast, the D+G strategy (drainage + green area) maintains a more linear decline in performance. The combination of structural and green infrastructure balances performance loss, moderating the pace of decline as γ increases. This indicates a more consistent risk reduction effect over time.

The D+G+Tun30 strategy shows a two-phase pattern: performance remains stable until around $\gamma = 0.25$, after which the RRR begins to decline linearly. This inflection point indicates that the addition of tunnel storage (Tun30) enhances early-stage robustness, delaying the onset of decline. Once γ exceeds the buffering capacity of the tunnel, the system transitions into a steady degradation phase.

We added a clarification in the text: "The differing curvature of the performance decline is due to the interaction of measures. For the Dr-only strategy, once implemented, its risk reduction gradually diminishes at an accelerating rate as climate stressors intensify – producing a concave-down curve (initially gentle slope, steepening later). In contrast, the strategies with multiple measures (D+G and D+G+Tun30) show a more linear decline. This is because when drainage alone begins to lose effectiveness, the next measure (GA, and later Tun30) either has just been implemented or is concurrently mitigating risk, effectively offsetting some of the non-linear drop. The combined result is a more steady (approximately linear) decrease in performance over time, as the measures' effects complement each other**.**"

Reference

Hu, H.Z, Tian, Z., Sun, L.X, et al. (2019). Synthesized trade-off analysis of flood control solutions under future deep uncertainty: An application to the central business district of Shanghai. Water research, 166, 115067.

RC2-17: *"Notation used in equations 1, 2 and 3 (L221-247) should be consistent with the notation adopted later in the Results section (L406-459)."*

Response: We completely agree that consistency in notation is critical. We have reviewed all equations (1, 2, 3) in the Methodology and cross-checked them against the notation in the Results section (which corresponds to original L406-459 and surrounding text).

In the revision, we unified the terminology and symbols. For example, if the Methodology introduced certain symbols for options, scenarios, thresholds, etc., we ensured the Results use the *same symbols* when referencing those concepts. Originally, we suspect there might have been a lapse, e.g., using "*a*" in the equations but then talking about "option" or "strategy" in results without the subscript, or using different letters like *x*, y later on. All such inconsistencies have been fixed.

RC2-18: *"Eq. 4 in the Supplementary material, Fig 3 in the manuscript: In the case of combined flood control measures implemented gradually (e.g., D+G, D+G+Tun30), do the authors consider the effects of the discount rate on the investment costs for additional interventions deferred over time? Are these effects reflected in the benefit-cost ratios plotted in Fig. 6?"*

Response: Thanks very much for this important issue. We realized we had not explicitly mentioned discounting in the manuscript. We have now stated it clearly:

(1) In our analysis, we did not incorporate a discount rate for future costs in the benefit-cost ratio (BCR) calculations – all costs were treated in constant terms for simplicity. We have now explicitly acknowledged this in the paper. For instance, in Section 2.2 or wherever we describe BCR, we added: "For simplicity, costs of measures were not discounted in our analysis, and we treated all monetary values in constant terms."

(2) We then addressed how this choice might affect results. Specifically, in Fig. 6 and the related discussion of BCR: We pointed out that if a strategy defers a major investment (like the Tun30) to later years, in reality its present value cost would be lower with discounting, potentially making the strategy more economically attractive than our simple BCR suggests. We added a sentence in the discussion of Fig.6: *(Discussion around Fig.6):* "**It should be noted that our BCR calculations do not incorporate discounting of future costs. In practice, implementing a tunnel later (as in D+G+Tun30) would incur discounted costs, likely improving the BCR relative to an early investment scenario. Including a standard discount rate would further favor pathways that defer expenditures; thus, our current BCR values are thus conservative for strategies with delayed measures.**"

(3) We also mention in the Conclusion or limitations that in future analyses, applying a discount rate would be more realistic and could change the economic ranking of pathways.

RC2-19: *"L432-450, Table 3; L615-616: the authors should provide the definition of density and coverage in the context of PRIM algorithm. Also, it is not clear how Table 3 is obtained by applying the condition given in Eq. 3."*

Response: We thank the reviewer for the helpful observation. *Coverage* is defined as the proportion of all scenarios that satisfy the identified condition (i.e., the scenario "box" coverage). *Density* is defined as the proportion of scenarios within that "box" that are successful (or meet the target criteria). We have added the missing explanations regarding the Patient Rule Induction Method (PRIM) results:

(1) In Section 2 L64, we add a brief sentence: "**In PRIM, 'coverage' refers to the fraction of the total scenario space included in the identified hyper-rectangle (subset), and 'density' refers to the fraction of cases in that subset that meet the specified success criteria**."

(2) In addition to the definitions of coverage and density, we now clarify how Table 3 is derived based on a combination of Equation (3) and the PRIM (Patient Rule Induction Method) algorithm.

Equation (3) defines the failure condition, where a strategy is considered to fail if its performance falls below the flood risk reduction target $F_0=0.7$ at any given level of the system stressor γ. PRIM is then applied to identify clusters of failure cases by searching across the full set of scenarios for each adaptation option. Specifically, for each option, we isolate the subset of scenarios that lead to failure according to Eq. (3).

Using the resulting failure boxes from PRIM, coverage is calculated as the proportion of all failure scenarios captured by the box, while density reflects the proportion of failure cases within the box relative to all scenarios it contains. Once these scenario clusters are identified, the Average Risk Reduction Rate (ARRR) of each option is calculated over its corresponding PRIM-derived subset.

In short, Table 3 integrates both the formal failure threshold logic (Eq. 3) and scenario discovery via PRIM to quantify the robustness of each option through three complementary metrics: coverage, density, and ARRR.

The manuscript has been revised to clarify this connection more explicitly for the reader in Section 4.2 from L433-L440: "Scenario discovery confirms that declining drainage capacity is the most critical uncertainty influencing the risk reduction rate. Failure scenarios are identified when the flood control target $F_0=0.7$ is not met. Using the PRIM algorithm, we extract failure-prone subspaces for each option by optimizing the combined value of coverage and density. Table 3 summarizes these metrics, where coverage and density are derived from PRIM-identified failure boxes, and ARRR is calculated as the average performance within those clusters. The valid period is defined by the point (characterized by γ) when a single option or combination no longer meets the performance target.”

RC2-20: *"L352: the authors state that they performed a trade-off optimization, balancing robustness and adaptiveness. However, later in the paper the authors clearly state that they actually did not run any optimization, which is deferred to future work (L637-641). I would clarify from the beginning that the contribution of the manuscript only focuses on the formulation and not on the solution of a specific complex optimization problem defined using the proposed framework. It is worth highlighting here that the case study considered in this work had a very small number of candidate solutions, which made it possible to "manually" enumerate and compare all of them, to identify the optimal solution; however, more realistic applications, with possibly hundreds of candidate flood control scenarios, will require using an optimization algorithm to reduce the number of hydrologic simulation that would be otherwise required if all scenarios were simulated. Incidentally, L639-640 include an incorrect statement, as genetic algorithms are not machine learning methods."*

Response: We appreciate this comment as it helps us clarify the scope and avoid confusion. We have made several changes to address these points:

(1) In the Introduction (around original L352) and again in the Methodology overview, we now clearly state that our work is about the *development of a decision-making framework* rather than executing a large-scale automated optimization. We explain that we formulated a method to consider both robustness and adaptiveness, and we demonstrated it by a case study with a manageable set of options. We explicitly mention that we did not run a computational multi-objective optimizer which would be typical if there were hundreds of options. Instead, we manually enumerated and effectively evaluated a small set of candidate solutions , given the case study’s scope. This sets the expectation correctly from the beginning.

(2) We added a sentence highlighting that the case study had a limited number of candidate strategies. Because of this small solution space, we were able to evaluate each and identify the optimal pathway without needing an algorithm to explore a huge design space. We then contrast this with a more complex real-world scenario: **we mention that if there were hundreds of possible interventions or sequences, then one would indeed need to integrate an optimization algorithm to search through the possibilities efficiently**.

(3) We went through instances where we did an optimization. For example, at original L352 we change phrasing “we performed a trade-off optimization” to “**we conducted a trade-off**

analysis". Similarly, at L637-641 (and 639-640), we ensure consistency. We emphasize that solving the optimization problem (especially via algorithm) was beyond the scope and is left for future work. This way, there is no contradiction: early on we admit we didn't do an algorithmic optimization, and later we reiterate that it's a future step for larger problems.

(4) We now refer to genetic algorithms appropriately as a heuristic optimization technique or an evolutionary algorithm. We removed the phrase "machine learning" in that context. We replace it with: **"...as genetic algorithms are an evolutionary optimization method**."

(5) In line with reviewer's note and also Comment #21, we took the limitations that were originally discussed late (L617-627) and created a dedicated subsection or earlier mention. We introduce a short "Limitations and future work" sub-section in the Methodology or Discussion where we say: "**this framework has certain limitations – one being that we only applied it to a small option set without formal optimization, and future studies should incorporate optimization for larger problems**." This ensures readers know these limitations *before* the conclusion.

RC2-21: *"L617-627: these limitations should be discussed earlier, possibly in a dedicated subsection of the revised Methodology section."*

Response: We have followed this advice. The limitations of our study (which were originally briefly listed around L617-627 in the conclusion/discussion) are now **discussed earlier** and more prominently in the manuscript.

After revising the Methodology and Results sections, we added a short subsection titled **"Limitations of the Framework"** (or a similarly appropriate title) either at the end of the Methodology section or at the beginning of the Discussion section – in any case, before the final conclusions. In this subsection, we outline the key limitations of our approach in a transparent way. These include, for example:

- The reliance on a simple hydrologic model (and the implications for generality and computational load, which we discussed in Comment #8).
- The small number of decision options considered (and that we did not employ a full optimization algorithm, as discussed in Comment #20).
- The simplified flexibility metric and related assumptions (as discussed in Comments #2-#4).

By aggregating these points in one subsection, we ensure that a reader can clearly see what the current study does not cover or where caution is needed in interpreting results. We then reference back to these limitations as needed.

RC2-22: *"L676-677: The code used to perform the analysis would be an important asset for the reviewers to evaluate the proposed decision-making framework and the case-study implementation, especially in light of the unclear methodology section. In line with HESS journal policies, I encourage the authors to share polished open-source code and the complete dataset to enable the reproducibility of their analysis."*

Response: Thanks for the suggestion. We agree that sharing our code and data will greatly enhance transparency and reproducibility. We are committed to open science and will make our materials available in accordance with HESS policies. We will prepare a replicable repository of our analysis code and a curated version of our dataset.

Minor comments:

RC2-23: *"L98-101 The flow in this sentence is broken, like if there were some missing connectors or else two individual sentences got messed up into one."*

Response: We have rewritten the sentence spanning Lines 98–101 to fix the broken flow. The original sentence was indeed awkward and seemed to combine two thoughts improperly. We split and restructured it for clarity. The corrected text in the Introduction now reads: "**However, an open question remains: is evaluating alternatives solely via robustness (with respect to decision-makers' risk aversion) sufficient to support rational long-term decisions? Or would a multi-objective trade-off analysis provide more comprehensive guidance?"**

This revision inserts the necessary connectors and separates the ideas into two coherent questions, addressing the reviewer's point about missing connectors.

RC2-24: *"L182 "the calculation of indicators such as the average risk reduction rate (ARRR) and benefit-cost ratio (BCR) for each alternative option was performed for each scenario". In this paragraph, some metrics are mentioned but their mathematical expression is not provided in the Methodology; additionally, the wording "indicators such as" is not appropriate in the Methodology section, as it is ambiguous. Does it mean that other metrics are calculated, besides ARRR and BCR? Or else do the authors intend that ARRR and BCR represent just examples of possible robustness metrics that the modeler may decide to use?"*

Response: No other hidden metrics were computed; ARRR and BCR were the main indicators of robustness in our analysis. The wording now unambiguously reflects that. We have clarified this part of the Methodology:

- We explicitly list the metrics calculated and provided their formulas in the Methodology section. For robustness evaluation, we calculated Average Risk Reduction Rate (ARRR) and Benefit-Cost Ratio (BCR) for each option under each scenario. In the revised text, we state this clearly and avoid the ambiguous "such as". We wrote: “For each candidate option and each scenario, we calculated two performance metrics: (1) the Average Risk Reduction Rate (ARRR) and (2) the Benefit-Cost Ratio (BCR).”

Additionally, we ensured ARRR and BCR are defined: e.g., ARRR is “percentage reduction in expected damages compared to baseline, averaged over scenarios” and BCR is “ratio of total avoided damage to total cost”. These definitions are now in the methodology so the reader knows exactly how they’re computed (connecting also to RC2-12 about performance metric).

RC2-25: *"L185 “Subsequently, the performance of each option and its combination was evaluated by quantitative comparison and ranking stability”. In this sentence, the notion of stability of the options is suddenly introduced, but it is never defined in the manuscript."*

Response: We realized that the term “ranking stability” was introduced without explanation. We have now defined what we meant by it:

In the revised Methodology, when we mention “ranking stability,” we added a brief definition. We intended it to mean how consistent the ranking of options is across different scenarios or evaluation criteria (i.e., does the “best” option remain best under slight changes, etc.).We clarified in text:“...evaluated by quantitative comparison, and the stability of each option’s rank across scenarios was assessed (i.e., how consistently an option remains superior or inferior under varying conditions).”

RC2-26: *"L188 “Valid periods of the alternative options were determined based on the conditions of the successful scenarios under each (individual or combined) option, in conjunction with a specific flood control objective.” From this sentence, the concept of “valid period”, quite redundant throughout the paper, is not clear (also see my main comment #9). A clearer description of such concept is given in the Results section (lines 438-439); however, it would be beneficial to the reader if the concept of “valid period” were introduced earlier, from the first time that it is mentioned."*

Response: As addressed in our response to RC2-9, we have replaced the term “valid period” with “**effectiveness period**” and, importantly, we now introduce and explain this concept at its first occurrence (which is around L188 in the Methodology).

In the revised Methodology text, we now say:"**We define the effectiveness period of an option as the duration (in years) for which that option remains effective in meeting the flood control objective. To determine this, we identified the point at which each option can no longer achieve the target performance (based on the conditions of the successful scenarios for that option)**."

This ensures that by the time the reader sees the results (like lines 438-439 in Results where the concept is used), they already know what it means.

We also cleaned up redundant mentions of "valid period" (now effectiveness period) so it's used consistently and only as needed.

Thus, we accepted and implemented the recommendation: the concept is clearly described upfront.

RC2-27: *"L200-205: concepts such as "transient scenarios" and "signposts" are unclear in this section of the methodology. They become clear to the reader after reaching the Results section (Fig. 6), but the article should be organized in a way that all concepts unroll smoothly following sections order."*

Response: We have added explanations for **"transient scenarios"** and **"signposts"** right when they are first mentioned in the Methodology (L200-205).

In the revised text, we explain:

- **Transient scenarios**: We clarify that by this we mean scenarios that change over time (e.g., evolving boundary conditions) as opposed to static scenarios. We likely reword "transient scenarios" to something more descriptive or add parentheses like "(time-evolving scenarios)".
- **Signposts**: We define signposts as decision-trigger indicators – measurable metrics or thresholds that signal when a change of strategy is needed.

For example, we now write: "We consider transient scenarios (scenarios in which conditions such as climate and land use evolve over time rather than remain static). In our adaptive pathway analysis, we also identify potential decision trigger points, or "signposts", which are predefined indicators (e.g., a certain flood frequency or performance threshold) that signal when an adaptation (policy shift) should be implemented."

By doing this, the reader is not left wondering what these terms mean. We agree that previously one would only infer their meaning later with Figure 6; now it's spelled out early. We made sure that the definitions here align with how they're used in Results (Fig.6 discussion), so that it's consistent throughout.

RC2-28: *"L208: "The choice of robustness option is the meta-problem of how to decide (Herman et al., 2005)". This sentence is too vague for and not related to the methodology section. I suggest either removing it or else moving it to the Introduction and elaborate more."*

Response: We found that this sentence was indeed out of place. It was a vague philosophical statement that didn't concretely help in the methodology. We have **removed this sentence** from the Methodology section to improve focus and clarity.

We did not find a strong need to elaborate it in the Introduction, as our Introduction already covers enough context. The removal did not affect the flow of ideas, and in fact made the Methodology cleaner.

RC2-29: *"L362-375: in this paragraph, parameters alpha, beta, and gamma are mentioned for the first time without prior introduction. These parameters are defined in the Supplementary material but should be also properly introduced in the manuscript."*

Response: Thanks for the reminder. We will introduce the parameters **α, β, γ** in the main text where they first appear.

RC2-30: *"L408-415: "Benefit-cost is the evaluation dimension for the robustness metrics" needs rewording; "benefit-cost, was defined as the average risk reduction rate (ARRR) per unit cost based on the robustness metrics of Laplace's Insufficient Reason" does not explain with sufficient detail how benefits and costs are calculated; also, the concept of Laplace's Insufficient Reason is mentioned a few times in the paper but never introduced."*

Response: We have rewritten the sentence at Lines 408-415 to clarify the meaning, and we have elaborated on how we calculate **benefit-cost** as well as introduced **Laplace's Principle of Insufficient Reason** if we invoke it:

- We rephrased "Benefit-cost is the evaluation dimension for the robustness metrics" to something more intelligible. Possibly we meant that we are evaluating robustness in terms of a benefit-cost ratio. We now say: "**We evaluate each strategy's robustness also in economic terms, using a benefit-cost ratio (BCR) as one of the criteria.**"
- We clearly explained how we compute **BCR**: We state what we consider "benefit" (likely risk reduction achieved, i.e., damage avoided) and what is "cost" (implementation cost of measures). We ensure the reader knows whether this is averaged over scenarios (Laplace's principle implies equal weights for scenarios) or worst-case, etc.
- We introduced **Laplace's Principle of Insufficient Reason** in the Methodology where we discuss robustness metrics. For example, when we mention regret or scenario

weighting, we add: "*(we assume all scenarios are equally likely, i.e., applying Laplace's principle of insufficient reason)*".

- Specifically in the Results around L408-415, we changed that confusing sentence to: "For an economic performance evaluation, we calculated a Benefit-Cost Ratio (BCR) for each option. The BCR was defined as the Average Risk Reduction Rate (ARRR, representing benefit in terms of proportional risk reduction) per unit cost of implementation. Here we applied Laplace's principle of insufficient reason – treating all considered scenarios as equally probable – when averaging outcomes for the BCR calculation."

In conclusion, we have carefully addressed all comments from Reviewer #2 by making the text more precise, adding needed definitions, reorganizing sections for better flow, and correcting any inaccuracies. We trust that these revisions have significantly improved the clarity and quality of the paper, and we thank the reviewer for his/her thorough and constructive feedback.