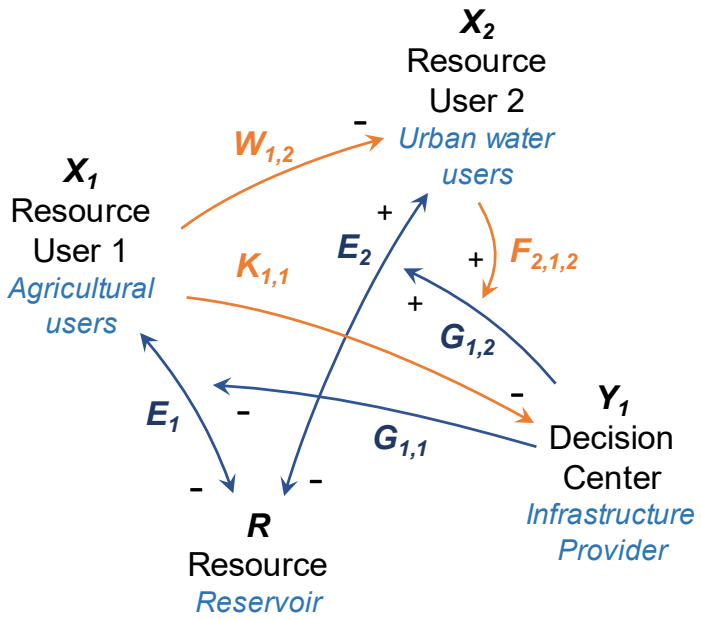
Response to Reviewer 1

This manuscript provides a very interesting and innovative set of stylized modeling experiments to explore the stability of governance structures in environmental systems. The evaluation of system stability across thousands of governance structures using a generalized dynamic systems modeling approach is particularly novel and insightful. The manuscript is well written and organized. I'd suggest the authors address the following comments to further improve the manuscript:

1. *The modeling approach necessarily deals with a stylized, abstracted representation of environmental governance systems. While there is some attempt to draw analogies between the mathematical abstraction and real-world systems in the introductory text, such analogies are largely excluded from the description of the model itself, the results, and the discussion/conclusion. I think the manuscript could be improved by providing examples of tangible aspects of real-world systems that the mathematical abstractions might represent (or using a single example, e.g., groundwater systems, and carrying it through the entire manuscript to aid readers with interpretability and bring the modeling formulation to life a bit more)*

We have revised Figure 1 and the caption to provide a more concrete example system that is referred to throughout the paper:

Example Parameter	Type	Meaning
$\frac{\sum_n E_n}{R}$	Scale	Extraction rate of resource (i.e. inverse of characteristic time scale of resource)
$\frac{E_1}{E_1 + E_2}$	Scale	Share of total extraction extracted by Resource User 1
$\frac{\partial E_1}{\partial R}$	Exponent	Sensitivity of Resource User 1's extraction to resource level



Example System Diagram. The nodes (R, X_1, X_2, and Y_1) are the state variables in the model, while the linkages represent functions (in blue) or parameters (orange) describing how the

variables interact. In this example water governance system, there are two types of water users, agricultural users and urban users, withdrawing water from a reservoir. The governance intervention $G_{1,1}$ in this example can be interpreted as infrastructure managed by the infrastructure provider, or Decision Center, that delivers water to the city, supporting urban extraction while reducing agricultural extraction. The orange linkages represent possible Nash Equilibrium strategies that may result from this setup. In this example, urban users allocate their effort to supporting the infrastructure that allows for their extraction ($F_{2,1,2}$), while agricultural users split their effort between undermining the organizational capacity of urban users ($W_{1,2}$) and of the Decision Center ($K_{1,1}$).

We have added the following references to this example system throughout the Modeling Approach section:

> In the example system, S represents the natural net gain to the reservoir after natural inflows and outflows that are not delivered to any users, and E_1 the total amount that agricultural users are able to extract.
>
> In Figure 1, $F_{1,2}$ is an example of such an effort that could represent urban users advocating for increasing the conveyance efficiency of the infrastructure delivering their water.
>
> These efforts are represented by $K^{+}_{k,m}X_k$ for supporting a venue, and $K^{-}_{k,m}X_k$, to undermine a venue. $K_{1,1}$ in Figure 1 for example, may represent agricultural users' efforts to undermine the authority of the Infrastructure Provider to withdraw water to deliver to urban water users.
>
> In Figure 1, for example, farmers divide their effort between undermining urban users' capacity ($W_{1,2}$) and undermining the capacity of the infrastructure provider that conveys water to urban users and away from farmers ($K_{1,1}$).

We have also added the following to the Results and Discussion section aid interpretation of the results:

> However, the results suggest that a greater effort put toward influencing the capacity of decision centers, or venue shopping, corresponds with stability, while greater effort put into the other strategies corresponds with reduced stability. In the example system, agricultural users are engaging in venue shopping by reducing the infrastructure provider's influence over the infrastructure ($K_{1,1}$); if there were other decision centers in the system, they may try to move that authority to a venue favors agricultural interests.
>
> Differences in this parameter correspond to different relationships with resource use: actors with low resource requirements, particularly if they are not involved in a profit-driven activity, may experience the largest capacity gains when their ability to extract is low. In contrast, some actors may become more invested and gain greater resources with which to mobilize as their extraction increases. In the example system, for example, urban interests will likely become less engaged once they have sufficient access to water (an inverse relationship between capacity and resource access), whereas agricultural users, particularly those part of industrial agriculture operations, might become less engaged once the available water, and thus profitability of farming, drops below a certain threshold.

2. *I'm not seeing where "stability" is clearly defined, both conceptually and mathematically. Perhaps I missed it. Regardless, a concise definition of the concept should be up front and center given the manuscript's focus.*

We have made the following changes to clarify the conceptual and mathematical meaning of stability and make those definitions more prominent:

In the introduction, we have moved the definition of stability earlier and added the mathematical definition in addition to the conceptual definition:

> Given that constant change is a central feature of complex systems, a system-level outcome of particular interest is stability. Mathematically, a steady state with local asymptotic stability is one for which trajectories near the steady state will approach the steady state. Conceptually, local asymptotic stability, hereafter referred to as stability, is an indication of the system's ability to retain its structure and function in the face of local perturbations in the variables controlled by the governance system (Guckenheimer and Holmes, 1983).

In the Modeling Approach section, we have revised the "Generalized Modeling Approach" heading to "Generalized Modeling Approach to Computing Stability" and have added the following text:

> Once the Jacobian is parameterized, the stability can be determined by checking whether the real part of all eigenvalues is negative. Conceptually, this means that perturbations in the state variables close to the steady state will return to that steady state. Local stability therefore indicates that the system will return to a steady state under short-term shocks (e.g. a sudden change to an actor's political influence), but does not necessarily indicate how the system will respond to large perturbations from the steady state or long-term drivers that fundamentally change the system's functioning (e.g. altering how resource users benefit from or impact the resource).

3. *While the effort to deploy thousands of structural variants of the environmental system is impressive and laudable, it seems to me that the revealed system dynamics still may be subject to higher-level structural assumptions regarding the nature of actor interactions. For example, NGOs are not directly tied to the state of the resource, whereas one might argue that NGOs are inversely (loss term rather than gain term) related to resource state (e.g., the tendency for environmental NGOs to emerge/grow as a particular environmental resource degrades). Likewise, actors could be viewed as operating within a nested structure (e.g., individual resource users interested in preservation of a resource comprising an NGO). While I understand that such a stylized formulation cannot touch upon all of these elements, I think the advantages/disadvantages of the proposed formulation can be further interrogated in the discussion.*

It is true that even with the Generalized Modeling formulation, some structural assumptions are inevitable.

We have added the following discussion of these types of higher-level modeling assumptions in the Conclusion section:

> Additionally, even though the generalized modeling approach requires fewer assumptions than traditional dynamical systems analysis, there are still assumptions regarding the structure of interactions among different model components. For example, the change in capacity of non-government organizations and decision centers does not directly depend on the resource state, but rather is affected by the resource state only indirectly through its influence on resource users' capacities and actions. Ultimately, we aimed to achieve a balance between a more general model that would make few assumptions about the structure of interactions, but would be challenging to interpret in the context of resource governance systems, and a more structured model, which limits

the variety of ways in which variables are linked but provides more precise insight into governance dynamics.

In the NGO case, the assumption is that if NGOs grow or decline based on the state of an environmental resource, it is due to support/lack of support from resource users in the system. This does not account for the fact that NGOs may have more trouble obtaining external sources such as grants or donations or may have more trouble recruiting staff when a resource is no longer threatened and vice versa since the assumption is that external forces will be less reactionary than resource users that are directly impacted by a resource. In a similar vein, regarding actors operating within a nested structure, while the model does not explicitly represent individual resource users comprising an NGO, it would represent a group of resource users collectively working with an NGO as a collaborative relationship in the model, where resource users help the NGO grow in capacity and their support depends on the state of the resource.

4. *Given the modeling interest on actors' ability to influence policies or capacities of other actors, there might be some interesting and relevant connections with the power relations and sustainability transitions literature (see for example Avelino and Wittmayer, 2016). Perhaps this could be further explored in the introduction and/or discussion/conslusion.*

Thank you for the suggestion, the power relations and sustainability transitions literature has clear connections with the concepts underlying the model. The following references have been added to the Modeling Approach section:

> This bottom-up perspective is chosen because of the under-representation of actors' agency in making and influencing decisions and pursuing their goals in the polycentric governance literature, which tends to focus solely on structure and exclude entities that lack the authority to create policies, though this is changing with concepts like institutional navigation (Dobbin, 2021; Villamayor-Tomas and García-López, 2018) and the sustainability transitions literature, which emphasize actors and the dynamic power relations among them as a driving force behind governance transitions (Avelino and Wittmayer, 2016}.

And to the conclusion:

> Additionally, while this study focuses on analyzing theoretical systems, the ability to model the different ways that actors exercise power and the dynamic power relations among them allows for exploring questions relating to the interaction between governance transitions and power relations in empirical systems as well (Avelino and Wittmayer, 2016; Avelino, 2021). This study demonstrates a way forward in combining the insights of complex systems theory with theories on governance to managing complex and highly uncertain human-natural systems in the face of rapid social and environmental change.

5. *The assumption of a Nash-equilibrium in actors' allocation of efforts is a strong one and receives very limited treatment in the manuscript. While I understand the adoption of the approach from a computational and conceptual standpoint, I think further elucidation of the implications and limitations of such an approach is warranted.*

This is a good point. Given that the strategy space is not necessarily convex, there is no guarantee of a Nash equilibrium. However, the modeling approach does not rely on the existence of a Nash equilibrium. To give results that are meaningful for understanding governance systems, the model only requires that actors behave in ways that are feasible (i.e. actions that don't contradict themselves) and are in their self-

interest, which the optimization method does ensure. We have added the following to the Modeling Approach section to clarify this point:

> A Nash equilibrium is calculated by computing the gradient of the equilibrium extraction or resource access and performing iterative steps of gradient descent for each actor in turn until the strategies converge. While there is no guarantee of a Nash equilibrium since the strategy space is not necessarily convex, the strategy optimization process ensures that even if optimality is not reached, actors are behaving in ways that are self-consistent and compatible with their goal of increasing their resource access. Modeling actors as behaving reasonably, if not necessarily rationally, ensures that the systems that are analyzed are feasible governance systems.

In addition, to address the concern as to whether a Nash equilibrium is a realistic representation of actor's strategies, we have added the following discussion to the conclusion:

> Additionally, the model assumes a Nash equilibrium in actors' strategies, representing actors as rational and having perfect knowledge of the system and others' actions, rather than the often heuristic and myopic manner in which they actually form their strategies for navigating governance (Pralle, 2003). However, this assumption is more reasonable in stable systems, where repeated interactions in a stable environment allow actors' greater opportunity to learn about the system and fine-tune their strategies (Craig et al., 2017; Pahl-Wostl, 2009).

6. *The abstract mentions a system's ability to "adapt to social and environmental change" and recover from "perturbations". Can the authors speak more to how perturbations of the system (in the form of either short-term shocks or gradual stressors factors) relates to the formulation? What exactly are the "perturbations in the variables controlled by the governance system" in this particular setup? And how does the concept of stability connect? I think a clearer defintion of stability (see comment above) and some added discussion could bring clarity to this.*

The paper uses the concept of a system's ability to recover from perturbations interchangeably with stability. In this setup, the "perturbations in the variables controlled by the governance system" refers to perturbations in variables such as actors' influences or resource state (i.e. state variables) as opposed to perturbations to parameters such as resource regeneration rates. The revisions in response to (2) above clarifies the connection between stability and system recovery to perturbations, as well as the types of perturbations that can be understood through local stability analysis.

7. *Minor editorial comments:*

   Figure 1 - mismatch between F2,1,1 in the legend and F2,1,2 on the figure

   Fixed.

   Line 126 - "non-government" to "Non-government"

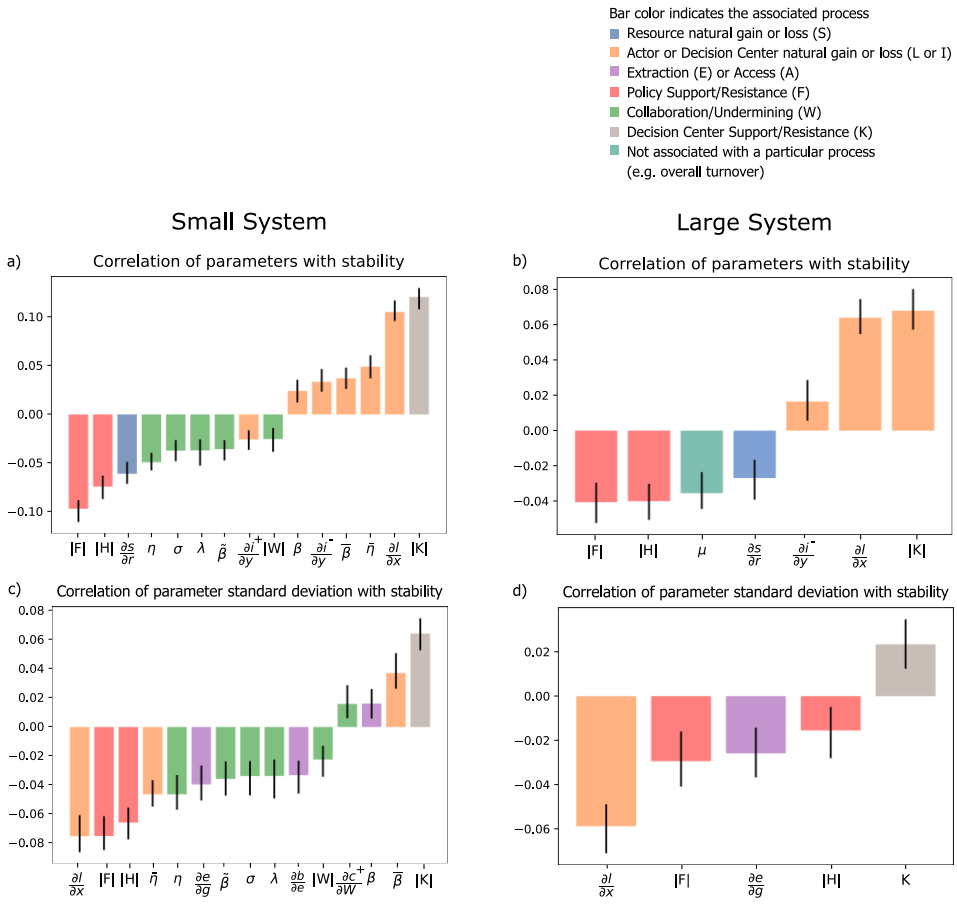
   Fixed.

   Line 138 - "them These" to "them. These..."

Fixed.

Figure 2 - Add "small system" and "large system" labels to the graphs (not just the captions) for readability

Figure 2 has been revised as suggested:



Response to Reviewer 2

This manuscript is a very welcome interdisciplinary contribution to Earth System Dynamics at both methodological and applied levels. Methodologically, it brings out solid dynamical systems approaches to addressing the highly nontrivial problem of environmental governance, where natural and human processes and interactions come into play that require not only the traditional dynamical systems principles in a sterile manner, but also social systems thinking with active decision making rather than the classical determinism. In this regard, this is a very insightful contribution that finds good grounds in an emerging but already reliable literature at the interface between natural and social systems with robust analytical mechanics principles and metrics (and dynamical systems in particular).

The stylised nature of the mathematical conceptualisations and experiments is crucial to shed light onto key interactions, with neither aiming at too much detail, nor at a too-macro of a picture that would wash out critical nonlinearities. As such, this is a very well balanced study, obviously with the inherent limitations that come with such exercise. The authors have done a pretty good job in laying

down their reasoning so that it is clearly understood where things come from and what they are meant to represent.

However, it is important to further clarify to those readership that is perhaps not so familar with one of either dynamical systems or governance reasoning the key notions being applied since aspects such as stability per se mean different things to different scientific communities. Further mathematical detail, while often discouraged in other venues, is never too much in this study, hence the authors are encouraged to add, perhaps in annex not to break the pleasant and clear flow of the text, further details on the underlying mathematical physics principles supporting their formal reasoning and formulation.

In order to clarify the meaning of stability, we have made the following revisions:

In the introduction, we have moved the definition of stability earlier and added the mathematical definition in addition to the conceptual definition:

> Given that constant change is a central feature of complex systems, a system-level outcome of particular interest is stability. Mathematically, a steady state that has local asymptotic stability is one for which trajectories near the steady state will approach the steady state. Conceptually, local asymptotic stability, hereinafter referred to as just stability, is an indication of the system's ability to retain its structure and function in the face of local perturbations in the variables controlled by the governance system (Guckenheimer and Holmes, 1983).

In the Modeling Approach section, we have revised the "Generalized Modeling Approach" heading to "Generalized Modeling Approach to Computing Stability" and have added the following text:

> Once the Jacobian is parameterized, the stability can be determined by checking whether the real part of all eigenvalues is less than 0. Conceptually, this means that perturbations in the state variables close to the steady state will return to that steady state. It is worth noting that local stability therefore indicates that the system will return to a steady state under short-term shocks to the steady state (e.g. a sudden change to an actor's political influence), but does not necessarily indicate how the system will respond to large perturbations from the steady state or long-term drivers that fundamentally change the system's functioning (e.g. altering how resource users benefit from or impact the resource).

While a full justification of the validity of the Generalized Modeling method is outside the scope of this paper, and we refer readers to Gross and Feudel, 2006 for this, we have expanded the Supplementary Information to include the full mathematical derivation of the generalized parameters and Jacobian, and the calculation of the objective function gradient. Please find the revised supplementary information attached.

The conditions under which their formulations are applicable and not should also be further discussed with additional few sentences so that the more naive reader is not tempted to throw the models around without enough care. The authors were clearly careful and that is very well seen through the solidity of their argumentation, formulation, results and discussion. But an additional pedagogic little touch would be the cherry on top of the cake to further help the increasingly mathematically fragile geoscience readership and even more so those coming from the more social science side that might alsot be interested.

We have made the following revisions to address the assumptions underlying the modeling approach:

We have added the following discussion of higher-level modeling assumptions in the Conclusion section:

> Additionally, even though the generalized modeling approach allows for making fewer assumptions than traditional dynamical systems analysis, there are still assumptions regarding the structure of interactions among different model components. For example, the change in capacity of non-government organizations and decision centers is modeled as not directly depending on the resource state, but rather being affected by the resource state only indirectly through how it changes resource users' capacities and actions. Ultimately, we aimed to achieve a balance between a more general model in which every variable can impact every other variable, making few assumptions about the structure of interactions but also limiting the insight into the dynamics specific to resource governance systems, and a more structured model, which limits the variety of ways in which variables are linked in the model.

We have added the following discussion about the assumptions underlying the search for a Nash Equilibrium to the Modeling Approach section:

> A Nash equilibrium is calculated by computing the gradient of the equilibrium extraction or resource access and performing iterative steps of gradient descent for each actor in turn until the strategies converge. While there is no guarantee of a Nash equilibrium since the strategy space is not necessarily convex, the strategy optimization process ensures that even if optimality is not reached, actors are behaving in ways that are self-consistent and compatible with their goal of increasing their resource access. Modeling actors as behaving reasonably, if not necessarily rationally, ensures that the systems that are analyzed are feasible governance systems.

In addition, to address the concern as to whether a Nash equilibrium is a realistic representation of actor's strategies, we have added the following discussion to the Conclusion:

> Additionally, the model assumes a Nash equilibrium in actors' strategies, representing actors as being rational and having perfect knowledge of the system and other's actions, rather than the often heuristic and myopic manner in which they actually form their strategies for navigating governance (Pralle, 2003}. However, this assumption is more reasonable in stable systems, where repeated interactions in a stable environment allow actors' greater opportunity to learn about the system and fine-tune their strategies (Craig et al., 2017; Pahl-Wostl, 2009).

Finally, to aid understanding of the abstract concepts in the model, we have added a concrete example system that is referred to throughout the Modeling Approach and Results and Discussion sections (see our response to Reviewer 1).

Last but not least, the remarks raised by the other referee are also hereby endorsed and will not be repeated. I would not say better in such regards.

All in all, this manuscript is definitely suitable for publication at Earth System Dynamics, is mostly appropriate at the scientific and technical levels safe for the minor aspects raised by both of us, and would also benefit from providing an extra layer of clarification and caution so that a broader readership other than us more technically minded can actually appreciate better the value and harness the vast potential of this contribution.

Thank you and all the best.