Does High-Resolution Downscaling Improve the Accuracy of Global Flood Model Inundation Estimates? An Analysis Across Biomes

A thesis submitted in fulfillment of the requirements
for the degree of

Master of Science
Ecology, Evolution, and Conservation

by
Julius Kindsgrab

University of Potsdam

Submission Date: 14.03.2025

Julius Kindsgrab, 806336
kindsgrab@uni-potsdam.de
Siemensstr. 13, 14482 Potsdam

Primary Supervisor:
Dr. Jacob Schewe (PIK)
Secondary Supervisor:
Prof. Dr. Oliver Korup (UP)

Declaration of Authorship

I hereby certify that the thesis I am submitting is entirely my own original work except where otherwise indicated. I am aware of the University's regulations concerning plagiarism, including those regulations concerning disciplinary actions that may result from plagiarism. Any use of the works by any other author, in any form, is properly acknowledged at their point of use.

14.03.2025,

Date, Signature

Acknowledgments

I would like to express my gratitude to my primary supervisor, Dr. Jacob Schewe, whose insightful guidance, support, and critique have been instrumental in shaping this research. I am also indebted to Professor Oliver Korup, for his valuable suggestions and for challenging me to enhance the depth and quality of my work. I sincerely appreciate Sandra Zimmermann for introducing me to the necessary tools and assisting me during the development of the experimental design. I gratefully acknowledge the invaluable institutional resources offered by the Potsdam Institute for Climate Impact Research.

I want to thank my friends for putting up with me and for making my life more fun, easy, satisfying and meaningful. Without the conversations, shared evenings, and trips taken, I would not have finished this thesis. I am especially indebted to Ernie, Nena, Philipp and Eric.

I am forever and immeasurably thankful for my family, without whom I would be lost. To my parents, Michael and Bettina, for always believing in me, even when I don't. And to my brothers, Paul and Steffen, for hilarity and depth, and making me want to be a better person.

I used LLMs through the provider ChatGPT as a research companion for several tasks: in finding relevant peer-reviewed literature from academic journals; in writing code; and in finetuning and clarifying some parts of the thesis text. ChatGPT provided me with the initial draft for much of my code, which I subsequently manually adjusted and further debugged in an iterative, cooperative process with ChatGPT. Approximately 5% of the text of my thesis was initially provided by an LLM, for instance through the summary of a research paper, which I however in every case subsequently manually adjusted to fit the style and logical flow of my thesis, and double checked for accuracy. The scientific writing tool Writefull further assisted in enhancing the clarity and style of my thesis, when I was at a dead end. Approximately 10% of my thesis was thus enhanced by Writefull.

Contents

Contents i

List of Abbreviations ii

List of Figures iv

List of Tables viii

1 Introduction 2
- 1.1 Background 2
- 1.2 Risk: Hazard—Exposure—Vulnerability 3
- 1.3 Flood Hazard 5
- 1.4 Biomes 6
- 1.5 Global Flood Models 7
- 1.6 Problem Statement and Thesis Goal 10

2 Methodology 12
- 2.1 Models 12
- 2.2 Observational Data 13
- 2.3 Case Study—Choice of Regions 14
- 2.4 Analysis 20

3 Results 23
- 3.1 Impact of Diagnostic Downscaling 23
- 3.2 Impact of Resolution-change Methodology 27
- 3.3 Performance Across Biomes 32

4 Discussion 35
- 4.1 Key Results 35
- 4.2 Diagnostic Downscaling Improves Performance and Validation Procedures Matter 36
- 4.3 Performance Differs Across Biomes 43
- 4.4 Conclusion 46

4.5 Outlook 47

Bibliography **48**

List of Abbreviations

AOM	Atlas of Mortality and Economic Losses from Weather, Climate and Water Extremes
CaMa-Flood	Catchment-Based Macro-scale Floodplain Model
CSI	Critical Success Index
DEM	Digital Elevation Model
DFO	Dartmouth Flood Observatory
GFD	Global Flood Database
GFM	Global Flood Model
GHM	Global Hydrological Model
ISIMIP	Inter-sectoral Impact Model Intercomparison Project
IUCN	International Union for Conservation of Nature
MODIS	Moderate-resolution Imaging Spectroradiometer
RCP	Representative Concentration Pathway

List of Figures

1 **Information on Weather Related Natural Disasters:** Distribution of (a) number of disasters, (b) number of deaths, and (c) economic losses by hazard type by decade globally. Reproduced from WMO [2021] 4

2 **An Overview of the Investigation Workflow:** The investigation workflow of observation and modeling, which meet in calculating performance scores. The analysis begins with a choice of events and subsequent generation of model flood discharge for these events. The resampling of observational data, sourced from the Global Flood Database, is tested with three resampling algorithms. Simulated flood extents are diagnostically downscaled to 1 arcmin and 15 arcsec, while observational data are resampled from 250 m using nearest-neighbor resampling. Simulated 1 arcmin is resampled to 15 arcsec, and simulated 15 arcsec is resampled to 250 m. Three analysis are executed based on three performance metrics. Events are further categorized into biomes. 15

3 **Illustration of Need for a Region-Subset:** On the left, the full-scale map derived from the Global Flood Database is shown, with flood-affected areas rendered in light blue and regions with missing data represented in gray. The black-outlined frame defines the region selected for further analysis. On the right, an enlarged view of the framed subset is provided. 16

4 **Global Biomes Map with Region Bounding Boxes:** Antarctica and Greenland are not mapped as biomes. 16

5 **Biome Maps of 6 Selected Regions:** The black-outlined frame defines the regions selected for further analysis. 19

6 **Consequences of 3 Resampling Algorithms:** Flood maps for one flood event; at its native 250 m resolution (top row), resampled to 15 arcsec (middle row) and resampled to 1 arcmin (bottom row). Each column represents one resampling method. In the any-presence resampling flood maps of 1arcmin and 15arcsec, white space denotes no-flood and non-catchment, and gray areas represent observed water. 20

7 **Performance Metrics at 1 arcmin and 15 arcsec Resolution:** Boxplots display the distribution of performance scores—CSI (1 indicates perfect model fit), Bias (0 indicates no bias, positive values overestimation, negative values underestimation), and Hit Rate (1 indicates entire flood captured)—for 18 simulated flood events diagnostically downscaled to two spatial resolutions. Orange line, box, and whiskers represent median, interquartile range (IQR), and maximum or minium values, respectively; numbers indicate minimum, median, and maximum values. 23

8 **Differences in Performance Metrics Between 1 arcmin and 15 arcsec Resolution:** Boxplots display the distribution of differences in performance metrics—CSI, Bias, and Hit Rate—between 18 simulated flood events diagnostically downscaled to 1 arcmin and 15 arcsec resolutions. For each flood event and metric, the difference (Δ Scores) was computed by subtracting the 1 arcmin score from the 15 arcsec score. The gray line at zero represents zero difference, positive values denote increase in scores, negative values a decrease. Gray line, box, and whiskers represent median, IQR, and maximum or minium values, respectively. 25

9 **Flood Maps for Three Flood Events at Resolutions of 1 arcmin and 15 arcsec:** Maps depict regions of concordance and discordance between simulated and observed water extents. **a)** Event ID 2279, Pakistan, (Biome: Desert); **b)** Event ID 3267, Bolivia (Biome: Tropical savanna); **c)** Event ID 2296, Northeast China (Biome: Flooded grasslands). 26

10 **Performance Metrics Across Three Resolutions and Two Resolution-change Methods:** Boxplots display the distribution of CSI (1 indicates perfect model fit), Bias (0 indicates no bias, positive values overestimation, negative values underestimation), and Hit Rate (1 indicates entire flood captured) scores for 18 simulated flood events across three resolutions and two resolution-change methods: Simulations diagnostically downscaled to 1 arcmin and 15 arcsec; simulation diagnostically downscaled to 1 arcmin and resampled to 15 arcsec (15sec_from_1min); simulation diagnostically downscaled to 15 arcsec and resampled to 250 m (250m_from_15sec). Orange line, box, and whiskers represent median, interquartile range (IQR), and minimum or maximum values, respectively; numbers indicate minimum, median, and maximum values. (Δ Scores) 27

11 **Differences in Performance Metrics Across Three Resolutions and Two Resolution-change Methods:** Boxplots display the distribution of differences in performance metrics—CSI, Bias, and Hit Rate—between 18 simulated flood events across three resolutions and two resolution-change methods: Simulations diagnostically downscaled to 1 arcmin and 15 arcsec; simulation diagnostically downscaled to 1 arcmin and resampled to 15 arcsec (15sec_from_1min); simulation diagnostically downscaled to 15 arcsec and resampled to 250 m (250m_from_15sec). For each flood event and metric, the difference (Δ Scores) was computed by subtracting the scores of the second part of the versus equation in the legend from the first part. The gray line at zero represents zero difference, positive values denote increase in scores, negative values a decrease. gray line, box, and whiskers represent median, interquartile range (IQR), and maximum or minium values, respectively. The x-axis marks the comparison method. 29

12 **Flood Maps of a Single Flood Event for Three Resolutions and Two Resolution-change Methods:** Maps depict regions of concordance and discordance between simulated and observed water extents. Resolutions and resolution-change methods are: simulations diagnostically downscaled to 1 arcmin and 15 arcsec; simulations diagnostically downscaled to 1 arcmin and resampled to 15 arcsec (15secfrom1min); simulations diagnostically downscaled to 15 arcsec and resampled to 250 m (250mfrom15sec). EventID: 2507, Myanmar, biome: Tropical rainforest. 30

13 **Performance Metrics at Two Resolutions Stratified by Biome:** Boxplots display the distribution of scores—CSI (1 indicates perfect model fit), Bias (0 indicates no bias, positive values overestimation, negative values underestimation), and Hit Rate (1 indicates entire flood captured)—between 18 simulated flood events. Simulations diagnostically downscaled to 1 arcmin (top row) and 15 arcsec (bottom row). Each event's deviation from the overall mean across all biomes is shown, stratified into biomes. Each panel represents one performance metric. The horizontal zero line marks no deviation from the overall mean score, positive values denote above average scores, negative values below average scores. gray line, box, and whiskers represent median, interquartile range (IQR), and maximum or minium values, respectively. 33

14 **Differences in Performance Metrics Between 1 arcmin and 15 arcsec Stratified by Biome:** Boxplots display the distribution of differences in performance metrics—CSI(1 indicates perfect model fit), Bias (0 indicates no bias, positive values overestimation, negative values underestimation),and Hit Rate (1 indicates entire flood captured)—between 18 simulated flood events diagnostically downscaled to 1 arcmin and 15 arcsec resolutions. For each flood event and metric, the difference (Δ Scores) was computed by subtracting the 1 arcmin score from the 15 arcsec score and subsequently stratified into biomes. The gray line at zero represents zero difference, positive values denote increase in scores, negative values a decrease. gray line, box, and whiskers represent median, interquartile range (IQR), and maximum or minium values, respectively. 34

List of Tables

1 **Basic Information on All Flood Events.** EventIDs are taken from the Global Flood Database. Bolded rows in the Biome column denote the biome assigned to subsequent analysis. 17

2 **Statistical Tests Across Three Resolutions and Two Resolution-change Methods and Three Performance Metrics:** Mean_Diff represents the average difference between the events at the two methods, while Std_Diff denotes the corresponding standard deviation. Normality of the differences was evaluated using the Shapiro-Wilk test, with both the test statistic and p-value reported (Shapiro-Wilk (stat) and Shapiro-Wilk (p), respectively). Based on the normality results, either a paired t-test or the Wilcoxon Signed-Rank test was applied (Chosen_Test). Test_stat and Test_p indicate the test statistic and the p-value, respectively, for the selected test. For every comparison, n=18 events were compared. 31

Abstract

Flood hazards are intensifying globally under changing climatic conditions, necessitating improved predictive capabilities in large-scale flood modeling. This study investigates whether high-resolution diagnostic downscaling of the CaMa-Flood global flood model enhances the accuracy of inundation estimates [Yamazaki et al., 2011]. Using a diagnostic downscaling approach, the model outputs originally generated at 0.1 ° resolution were refined to a 1 arcmin and 15 arcsec grid, and the resulting flood extents were validated against observational data from the Global Flood Database, resampled at those same resolutions using nearest neighbor resampling [Tellman et al., 2021]. Three performance metrics—Critical Success Index (CSI), Bias, and Hit Rate—were employed to quantitatively assess model accuracy. In a series of comparative analyses across six floodplains spanning four distinct biomes (tropical savannah, desert, flooded grasslands and tropical rainforest), diagnostic downscaling consistently improved CSI values and reduced overestimation bias at 15 arcsec relative to both the 1 arcmin simulations and those resampled from 1 arcmin to 15 arcsec via nearest-neighbor resampling. Furthermore, secondary analyzes revealed that the gains achieved at 15 arcsec relative to 1 arcmin were partially facilitated by a more substantial decrease in the fidelity of the observed dataset as it was further resampled from its native resolution. Tertiary investigations highlighted biome-specific variations in downscaling efficacy, suggesting that local ecological conditions might modulate model performance. These findings substantiate that high-resolution diagnostic downscaling is a promising avenue for reducing uncertainties in global flood modeling and holds potential for informing more effective flood risk management strategies under future climate scenarios. However, they also reveal the present limitations, suggesting areas for future enhancement and underscoring the challenges in unambiguously validating flood models.

1. Introduction

Floods have played a fundamental role in the formation of human civilization, as evidenced by the presence of flood myths in almost every culture. For instance, ancient Hindu texts recount how Manu, the first man, was warned of an impending deluge and survived it alone; the Epic of Gilgamesh describes the god Enlil unleashing a catastrophic flood to destroy the world; and the Old Testament narrates the story of Noah's Ark, a pivotal account in the Judaeo-Christian tradition [Leeming, 2022].

Although these stories are not about floods per se, but rather symbolize the "cleansing of a sinful world" [Leeming, 2022], the use of floods, as opposed to other natural disasters, does underscore the deeply rooted understanding, fascination and reverence humans have had for the immense, awe-inducing power of these events. This is only substantiated by the fact that these myths arose independently across diverse regions. In these narratives, floods are portrayed as forces that bring not just destruction but renewal, laying the foundation for life and the rise of civilization. This dual nature is also reflected in reality, where river floods have not only brought despair and reshaped continents, but also provided vital irrigation, thereby enabling life and the development of human societies [Leeming, 2022].

1.1 Background

Today, floods are one of several natural disaster types that is being changed and amplified by climate change. Data from the Atlas of Mortality and Economic Losses from Weather, Climate and Water Extremes (AOM) (1970–2019) [WMO, 2021] reveal that in this timeframe, more than 2 million deaths and US$3.6 trillion in economic losses were caused by weather, climate and water-related hazards. This accounts for 45% of all disaster-related deaths and 74% of economic losses worldwide during this timeframe [WMO, 2021]. Temporal analysis shows that the frequency of disaster events has increased, a trend influenced by both improved disaster reporting and the potential amplification of hydrological extremes due to climate change [UNDRR and CRED, 2020, WMO, 2021].

According to the AOM [WMO, 2021], storms were the leading cause of mortality and incurred the highest economic losses (figure 1). Fortunately, over the decades, mortality rates from natural disasters have significantly declined. Annual death tolls dropped from more than 50,000 in the 1970s to less than 20,000 in the 2010s, despite a doubling of the total global population. This decline is largely attributed to advancements in early warning systems and disaster preparedness. In contrast, economic losses escalated dramatically, increasing sevenfold from the 1970s to the 2010s. This increase was caused by and highlights the growing exposure and asset accumulation in hazard-prone areas [WMO, 2021]. The global impact of disasters exhibited stark regional disparities. Developing economies experienced the majority of deaths, with 91% of deaths occurring in these regions. Conversely, developed economies bore a disproportionate share of the economic losses, accounting for 59% of the total. However, although high-income countries incurred the largest absolute financial losses, this is due to their higher overall economic exposure to floods. In contrast, lower-income countries suffered the highest relative losses [UNDRR and CRED, 2020].

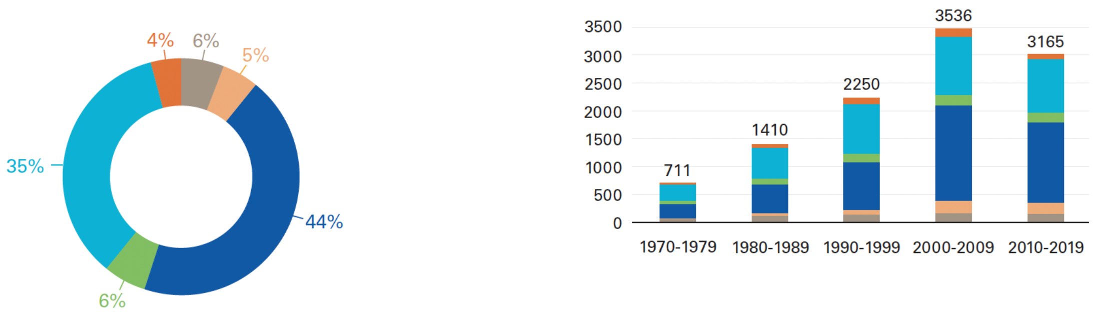
Riverine floods are the most frequently occurring weather-related disaster, accounting for 44% of all recorded disaster events worldwide between 1970 and 2019. Although floods were not the leading cause of mortality, they still caused tens of thousands of deaths, often concentrated in large-scale events. In addition, floods were responsible for 31% of the total economic losses caused by weather, climate, and water-related disasters [WMO, 2021].

1.2 Risk: Hazard—Exposure—Vulnerability

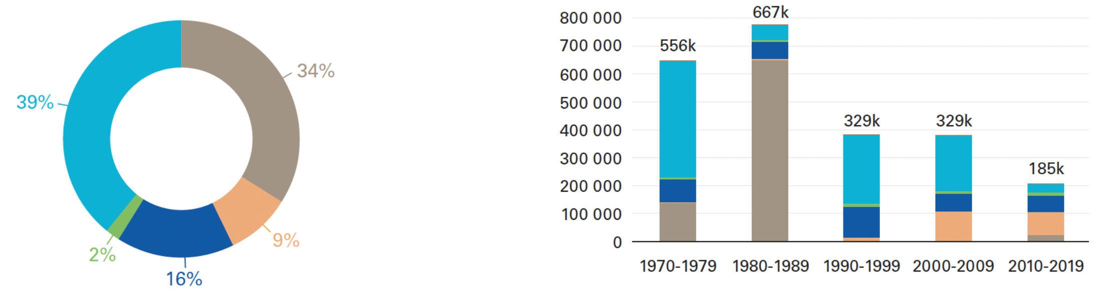
The risks posed by natural disasters in general and floods specifically can be conceptualized as a dynamic and interdependent interaction between three primary elements: hazard, exposure, and vulnerability [Koks et al., 2015, Reisinger et al., 2020]. A hazard is a potentially damaging physical event or phenomenon (for example, a flood) that can have adverse impacts on people, infrastructure, or the environment. Exposure refers to the presence of people, assets, infrastructure, and ecosystems in areas where they are susceptible to the impacts of hazards. Vulnerability is the propensity of exposed elements to suffer harm or loss due to hazards, influenced by socioeconomic, environmental, and structural factors that determine their resilience [Koks et al., 2015, Reisinger et al., 2020].

Risk in general can be viewed as a ‘contingent liability’ for ‘adverse consequences for human or ecological systems’, in which allowing future risks to accumulate weakens a nation’s potential for socioeconomic development [Ward et al., 2015, Reisinger et al., 2020].

(a) Number of reported disasters
Total = 11 072 disasters



(b) Number of reported deaths
Total = 2 064 929 deaths



(c) Reported economic losses in US$ billion
Total = US$ 3.6 trillion

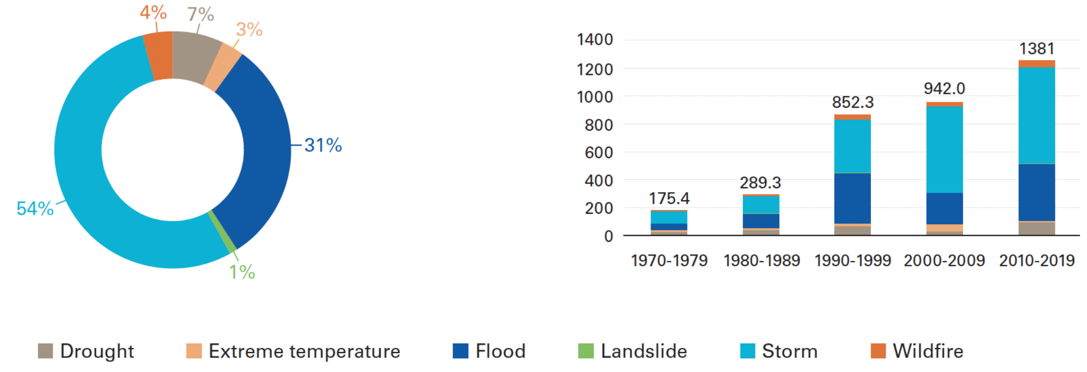


Figure 1: **Information on Weather Related Natural Disasters:** Distribution of (a) number of disasters, (b) number of deaths, and (c) economic losses by hazard type by decade globally. Reproduced from WMO [2021]

This makes managing the risks from these disasters essential, which in turn requires a deep understanding of risk. This includes understanding both the current state of each risk factor (hazard, exposure, vulnerability) and predicting the future development of each factor [Ward et al., 2020].

Uncertainty in Modeling and Broad Context

In both cases, understanding the current state and projecting into the future, there are technically solvable and theoretically irreducible issues that limit the certainty that can be achieved in risk quantification and thus risk management. Theoretically irreducible issues relate to the fundamental limitations of models in representing the full reality of the current state, as well as modeling the inherently unpredictable complex adaptive interactions of the social-ecological system. This affects both the physical hazard modeling and the modeling of exposure and vulnerability, as all are subject to changes in human behavior and technological innovation.

Technically solvable issues relate to understanding physical laws and their interdependencies that result in hydrological processes and formulating these in mathematical equations; increasing the availability of data on the current state of relevant parameters and increasing the availability of compute to run sophisticated models. In this pursuit, ‘parsimonious’ models, those with fewer parameters, should be preferred, as further complexity does not necessarily equate to improved performance or understanding. In many instances, added complexity can lead to diminishing returns, where marginal gains in predictive accuracy are outweighed by increased computational costs and potential for overfitting. Occam’s Razor generalizes this as a philosophical principle that advocates simplicity in the construction of explanatory models: ‘Entities should not be multiplied without necessity’, suggesting that, among competing models that explain a given phenomenon equally well, the simplest one should be selected.

Due to these fundamental and practical limitations, any model forecast needs to be prefaced by, at minimum, ‘all factors holding steady’, and should not be misunderstood as a prediction of the actual state of reality for the time of prediction. With this caveat in mind, any reduction in uncertainty of model outcomes is to be welcomed and striven towards. Thus, in the broadest sense, this study is situated in the pursuit of uncertainty reduction for the hazard part of the risk equation for river floods.

1.3 Flood Hazard

The literature on changes in river flood hazards falls into two general categories: those studying past observations and those examining future model-based projections. Both types of studies suffer from methodological constraints. Studies modeling future (or past) projections of river flood hazards face the aforementioned set of constraints. Studies focusing on past observations face limitation in the availability and coverage of data, with historical records often exhibiting spatio-temporal gaps. Observational networks are unevenly distributed, with high-quality data concentrated in developed regions and sparse coverage in remote and underdeveloped areas. Reporting biases may further influence the available flood records.

That being considered, the current literature on observed changes in flood hazard reveals a general increase in the frequency and magnitude of global floods and affected areas. However, significant spatial and temporal variability across different climate zones exist, reflecting the interplay of climatic changes, anthropogenic influences, and methodological constraints [Slater et al., 2021, Liu et al., 2022]. On a global scale, the frequency of floods has increased significantly between 1985 and 2015. Flood duration has also experienced notable changes, with the global median increasing from four days in 1985 to ten days in 2015. This shift, from predominantly short-duration floods to more frequent moderate-duration events, indicates a systemic change in flood dynamics. These trends suggest an intensification of hydrological extremes, particularly in regions with strong climatic oscillations [Najibi and Devineni, 2018]. The annual average area affected by floods has also increased over time, reflecting the growing intensity and geographical extent of flood events [Liu et al., 2022]. Overall, the literature suggests a redistribution of flood risks, with increasing hazards in wetter regions and decreasing risks in drier regions [Slater et al., 2021, Liu et al., 2022, Najibi and Devineni, 2018].

The literature on future projections of flood hazards generally demonstrates that both the magnitude and frequency of river floods are highly sensitive to climatic variations and show stark regional disparities [Arnell and Lloyd-Hughes, 2014]. Moreover, the connection between the increase in global average temperature and the risk of flooding indicates nonlinear dynamics. As temperatures increase, the risk of flooding initially increases slowly, but accelerates at higher levels of warming [Arnell and Gosling, 2016, Alfieri et al., 2017]. Do et al. [2020] find increased flood hazards under the Representative Concentration Pathway (RCP) 6.0, with up to 35% of the grid cells showing statistically significant increases in flood magnitude, compared to fewer than 12% under RCP2.6.

1.4 Biomes

Ecological processes exert a substantial influence on flood hazards, operating through two principal modes of ecosystem-based regulation. First, at the catchment scale, ecosystems provide what may be termed “flood prevention ecosystem services” [Vári et al., 2022]. In this context, vegetation plays a critical role by intercepting rainfall through canopy and litter layers, thereby significantly diminishing the volume of precipitation that directly reaches the ground. Simultaneously, root systems enhance soil infiltration, which serves to reduce both the quantity and velocity of surface runoff before flood waves can develop [Vári et al., 2022].

In contrast, once water has concentrated in rivers, the regulatory function shifts primarily to adjacent ecosystems, particularly floodplains. These areas contribute to flood mitigation ecosystem services by providing additional storage capacity, thereby lowering the peak intensity and overall severity of flood waves [Vári et al., 2022]. When floodwaters inundate floodplains, the increased surface roughness—attributable to both the terrain and riparian vegetation—serves to decelerate water flow, thereby altering the spatiotemporal evolution of flood waves and, consequently, the associated flood risk [Rajib et al., 2023].

One possible level at which to investigate the effect of ecology on flood hazard is the biome level. According to the “IUCN Global Ecosystem Typology” [Keith et al., 2020], a hierarchical classification system, biomes are defined as “components of a core or transitional realm united by one or a few common major ecological drivers that regulate major ecological functions”.

Biomes are thus “biotic communities” occurring at large geographic scales, determined primarily by climate, and recognizable by a typical vegetation physiognomy (overall structure and physical appearance of the dominant species within the community) [Cardoso et al., 2021, Keith et al., 2020, Mucina, 2019]. For example, a tropical rainforest biome is characterized by tall, evergreen broadleaf trees and high biomass, regardless of exactly which tree species are present. Biomes thus represent the global-scale patterns of ecosystems recurring under similar climatic conditions and are best delineated by combinations of climate and plant functional characteristics [Mucina, 2019, Keith et al., 2020, Cardoso et al., 2021]. However, the literature notes that biome definitions have varied historically, and there remains no single universally accepted method for delimiting biomes on Earth [Keith et al., 2020, Cardoso et al., 2021]. Despite these discussions on the definition and functional relevance of biomes, they are frequently used as tools to provide large-scale (regional to global) backgrounds in a range of ecological and biogeographical studies [Mucina, 2019, Keith et al., 2020].

1.5 Global Flood Models

Flood hazard is increasing and projected to increase further, thereby contributing to heightened flood risk and threatening socio-economic development. Fortunately, disaster risk management tools are available, mitigating flood risk through early warning systems and cost-effective prevention measures, such as dams, dykes, drainage systems, and increasingly, nature based solutions [UNDRR and CRED, 2020]. However, effective implementation of measures depends on identifying flood-prone areas based on return intervals, which makes flood models and proper observation essential for risk management and mitigation [Risling et al., 2024, Bernhofen et al., 2018].

Established practices in floodplain planning often rely on complex, locally calibrated hydrodynamic models [Ward et al., 2015]. However, these local models require detailed calibration for each catchment, depending on local observational data, which are not always readily available globally [Mester et al., 2021]. This is where Global Flood Models (GFM) offer promising potential. Although GFMs offer less manual intervention and limited integration of local knowledge, they provide comprehensive coverage and consistent methodologies across regions and are thus critical tools for large-scale flood prediction, especially for regions lacking local or national flood mapping programs [Wing et al., 2024, Ward et al., 2015, Risling et al., 2024, Bernhofen et al., 2018].

In recent years, advances in computing power, numerical algorithms, and remote sensing data accuracy have accelerated the development of GFM [Bernhofen et al., 2018, Risling et al., 2024]. GFMs are instrumental in predicting flood frequency, duration, intensity, and extent of affected areas [Bernhofen et al., 2018]. They facilitate and assist in answering practical and scientific questions, such as estimating the population or land area exposed to floods, the value of assets at risk, the costs and benefits of flood defenses, and the effects of climate and socio-economic changes on future flood losses [Hirabayashi et al., 2013, Ward et al., 2015, Alfieri et al., 2017, Dottori et al., 2018]. One of these models, the Catchment-Based Macro-scale Floodplain model (CaMa-Flood) by Yamazaki et al. [2011], is to be investigated in this thesis.

However, due to limitations in resolution, GFMs still fall short in providing the detailed, locally specific flood risk information required for designing individual flood defenses. In CaMa-Flood, diagnostic downscaling redistributes river discharge volumes over finer digital elevation models (DEM) to enhance the spatial resolution of flood simulations up to 3 arcsec (90 m at the equator). The method conserves mass and momentum, thus ensuring that water volumes are accurately represented at smaller scales, reflecting local variations in terrain [Yamazaki et al., 2011, Wing et al., 2024]. While higher resolutions hold the potential to enhance the accuracy and thereby practicality of GFMs, the necessary increase in model complexity needs to be legitimized by significant gains in model accuracy, and furthermore introduces challenges related to validation (data interpolation), computational demands and overfitting [Ward et al., 2015, Mester et al., 2021]. Any potential added benefit of diagnostic downscaling needs to be carefully weighed against its costs, by validating simulations at several resolutions and assessing the cost-benefit ratios of higher resolutions.

Validating Global Flood Models

Validating the accuracy and consistency of GFMs is difficult, especially where the availability of official flood hazard maps is limited. Challenges to comparison datasets include incompleteness, fragmentation, bias, and variations in reporting conventions [Risling et al., 2024, Bernhofen et al., 2018, Mester et al., 2021, Sampson et al., 2015, Ward et al., 2020]. Validation thus has often had to rely on indirect methods, such as cross-comparing model output from a set of GFM. These comparisons have been subject to marked inconsistencies and significant variability, showing merely a 30-40% agreement of flood extent [Risling et al., 2024].

Therefore, in recent years, there has been a significant increase in efforts to validate GFMs using satellite images of past flood events. This method involves a change detection analysis to identify changes in surface water extent before and during floods. Historical flood maps from databases such as the Dartmouth Flood Observatory (DFO) and the Global Flood Database (GFD) have been instrumental in these validation efforts. For example, Bernhofen et al. [2018] evaluated six GFMs by comparing their flood extent predictions with satellite observations from the DFO, finding agreement levels between 45% and 70%. Validating GFMs with openly accessible flood maps is therefore a crucial step toward improving their accuracy and performance [Risling et al., 2024].

For GFMs, previous papers have argued that ‘a pragmatic approach to [validation] is to compare the results against an analogous regional data set constructed using high-resolution models and detailed local data’ [Sampson et al., 2015]. This has been taken to mean that a pragmatic approach to validating the results of a GFM is to compare the results against high-resolution datasets [Risling et al., 2024]. However, the pragmatic part of the argument seems to have been in reference to the availability of the dataset, not the resolution.

In any case, validation requires both model output and observation output to have the same resolution, which in most cases, they do not, especially when multiple model resolutions are investigated. Therefore, either or both model output and observation output require a change in resolution. Enhancing the resolution of flood maps involves refining the spatial detail of either models or observations by integrating finer-scale information (diagnostic downscaling of CaMa-Flood) or resampling data to produce different-resolution outputs. In diagnostic downscaling, flood volume in a low resolution grid cell (0.1 °) is diagnostically distributed onto the underlying higher-resolution grid cells according to their elevation. Increasing the resolution through resampling amounts to an implicit assumption that the information content of the images will increase [Atkinson, 2013]. Bernhofen et al. [2018], in their validation study, desiring to preserve the detail of high-resolution models, resampled observed satellite imagery using the nearest-neighbor resampling method [Bernhofen et al., 2018]. Similarly, resolution can be decreased with resampling methods.

1.6 Problem Statement and Thesis Goal

To finally summarize and broadly contextualize the problem addressed by this study: Based on future projections and past analysis, climate change exacerbates the risk of climate-related natural disasters by increasing both the frequency and magnitude of hazards, a trend that extends to river floods. Flood risk, which undermines socio-economic development, is a dynamic and interdependent function of hazard, exposure, and vulnerability. This complexity poses significant challenges for accurately modeling and managing river flood risk, necessitating improvements across all three dimensions. The present study contributes to the improvement of hazard modeling by focusing on GFM, which are widely used tools capable of achieving high resolutions (approximately 90 m at the equator) by diagnostic downscaling. However, the efficacy and cost–benefit ratio of such downscaling methods remain unclear, underscoring the need for rigorous validation studies to inform GFM improvement and application. Validation requires a comparison dataset, often derived from remote sensing data, and, for meaningful analysis, both the modeled and observational datasets must be harmonized to the same spatial resolution through appropriate resampling techniques.

In this thesis, the added benefit to the accuracy of diagnostic downscaling in CaMa-Flood is systematically investigated, by downscaling CaMa-Flood simulations to two resolutions (1 arcmin and 15 arcsec) and validating the resulting flood maps to observations derived from the Global Flood Database [Tellman et al., 2021]. This database contains flood maps at a spatial resolution of 250 m, portraying the maximum surface water extent observed during large flood events.

To assess potential ecological impacts on modeling performance in general and diagnostic downscaling in particular, flood events from a diverse set of ecological contexts are chosen. The flood events investigated are large and contain multiple ecosystems within them. Therefore, the biome, as an ecological concept approximating this size, was chosen. Characterizations of biomes by their physiognomy and large-scale characteristics of soil and water availability lend themselves to an investigation of how the efficacy of flood modeling differs across varied ecological conditions.

The analysis will focus on a set of 6 floodplains distributed over 4 biomes and multiple flooding events per floodplain. For each event, the CaMa-Flood simulation will be diagnostically downscaled to 1 arcmin and 15 arcsec, whilst the GFD observations will be resampled to the same resolutions using nearest-neighbor resampling. For each event and comparison, three validation metrics will be calculated—Critical Success Index (CSI), Bias and Hit Rate. To further enhance clarity, flood maps for each resolution will be compared (primary analysis).

To investigate the effect of resampling in validation procedures, several sensitivity analyses will be performed (secondary analysis). To compute the benefit of diagnostic downscaling in CaMa-Flood versus downscaling by resampling, the difference between accuracy scores when diagnostically downscaling from 1 arcmin to 15 arcsec and resampling from 1 arcmin to 15 arcsec is calculated. Furthermore, the study will also resample the simulation from 15 arcsec to 250 m. Finally, a tertiary analysis will examine whether performance differences can be detected across biomes.

In summary:

1. **Primary Analysis**: The analysis will evaluate whether diagnostic downscaling of CaMa-Flood data to 1 arcmin vs. 15 arcsec is associated with measurable changes in performance metrics (CSI, Bias, Hit Rate).

2. **Secondary Analysis**: The study will compare diagnostic downscaling with nearest-neighbor resampling to separate the factor of spatial resolution from the process by which it is attained, and determine the effect of validation procedures.

3. **Tertiary Analysis**: The investigation will assess whether performance outcomes and the effects of diagnostic downscaling vary across different ecological contexts, as represented in biomes.

This study aims to determine the effect of increased model complexity (enhanced spatial resolution) on accuracy; how validation procedures influence performance; and whether ecological factors are relevant. The results will be useful for interpreting results from low-resolution global simulations and for guiding future choices for large-scale flood modeling in the context of climate impact assessments.

2. Methodology

2.1 Models

As a global flood model, I use CaMa-Flood (v. 4.0) developed by Yamazaki et al. [2011], a large-scale hydrodynamic model that simulates river discharge and floodplain inundation across global river networks. It is a cascade model that uses a precipitation time series from global climate reanalysis data to drive a global hydrological model, which produces flows across a river network [Bernhofen et al., 2018, Risling et al., 2024]. It operates at a native resolution of 0.1 ° and describes floodplain inundation dynamics by explicitly integrating subgrid-scale topography. In version 4.0 the model author has derived the underlying river network in CaMa-Flood based on MERIT-Hydro flow direction maps, using the MERIT digital elevation model (DEM) for its baseline topography [Yamazaki et al., 2019]. The subgrid-scale topographic parameters based on DEM of varying resolutions decide the relationship between water storage, water level, and the flooded area in the model. These fine-scale variations in topography are crucial for understanding how water spreads during flood events [Yamazaki et al., 2011].

The model runs were set up according to the Inter-sectoral Impact Model Intercomparison Project (ISIMIP) 3a simulation protocol [Frieler et al., 2024]. I use a single combination of climate forcing and global hydrological model (GHM). The climate forcing dataset used to drive the GHM is a combination of W5E5 v2.0 [Lange, 2019, Cucchi et al., 2020] with GSWP3 v1.09 [Kim, 2017, Dirmeyer et al., 2006]—GSWP3-W5E5. The dataset is derived from reanalysis products that incorporate data from regional weather stations and later adjust the precipitation values and additional variables using observational data from these stations. GSWP3-W5E5 further amends for precipitation undercatch by rain gauges [Mester et al., 2021].

The GHM used is WaterGAP2-2e [Müller Schmied et al., 2023]. The Global Assessment and Prognosis (WaterGAP) is a modeling technique designed to measure water resources and usage across every land region on Earth. The GHM runoff is used as input for CaMa-Flood v4.0, which provides discharge and measures flood depth on a grid with a resolution of 0.1 ° [Mester et al., 2021].

I then downscale this data to 1 arcmin and 15 arcsec. In downscaling, the daily flood volume in a low resolution grid cell (0.1 °) is diagnostically distributed onto the underlying medium- (1 arcmin) to high-resolution (15 arcsec) grid cells according to their elevation. Then, I assign to each high-resolution grid cell the maximum annual daily value, resulting in an annual flood depth time series. This method is taken from a previous study by [Mester et al., 2021], which works with imagery from the Dartmouth Flood Observatory (DFO), based on data from the Moderate-resolution Imaging Spectroradiometer (MODIS). In this study, “the event duration according to the satellite imagery [DFO] matches with the rising limb or the peak of the flood simulations for most regions of interest; and coincides with no second flood event in the year of investigation, which legitimizes this approach” [Mester et al., 2021].

2.2 Observational Data

I use observational flood data from the Global Flood Database (GFD), which is based on events documented in the DFO. The GFD was developed to systematically map the maximum observed surface water extent during 913 large floods between 2000 and 2018 [Tellman et al., 2021]. For the GFD flood maps, surface water is detected at a spatial resolution of 250 m and represented as either flooded or not flooded. This is done using data from MODIS, a multispectral optical instrument on NASA’s Terra and Aqua satellites, which effectively resolves large, slow-moving floods. The GFD flood maps incorporate permanent water data sourced from the JRC Global Surface Water Dataset, maintaining the initial 30 m resolution. This dataset characterizes permanent water using Landsat imagery collected from 1985-2016 [Pekel et al., 2016, Tellman et al., 2021].

Identifying flooded regions using MODIS encounters several constraints. Foremost, these relate to detecting floods under cloud cover and canopies, along with issues of cloud and mountain shadows [Sauer et al., 2024]. In the development of the GFD, each map passed a quality control process, ensuring that the flood plains presented in this study accurately depict the flood event [Sauer et al., 2024, Tellman et al., 2021]. Using a DEM, regions with slopes exceeding 5 ° are omitted from the final classification to minimize confusion with terrain shadows [Tellman et al., 2021]. In the qualitative flood-map comparison plots in this thesis, these areas are indicated as no-data and are thus gray.

Several uncertainties persist with the GFD. To mitigate pixel misclassifications, the GFD employs 3-day and 2-day composites, ensuring that a pixel is designated as a water pixel only if at least half of the observations during the specified time period identify it as water [Tellman et al., 2021]. In areas with rapid water movement, the temporal resolution of two daily images, coupled with possible cloud coverage, can result in underestimating the flood extent [Sauer et al., 2024].

Figure 2 presents an overview of the modeling and overall methodological process of this study.

2.3 Case Study—Choice of Regions

I chose 6 regions from 4 different biomes for the case study. For each region, several flood events are available and will be analyzed. The regions needed to meet the following conditions:

1. The regions need to be part of the GFD, as availability of observational validation data that is consistent and comprehensive throughout the entire region of interest is crucial [Mester et al., 2021].
2. To adequately capture inundated areas considering the spatial resolution of the GFM, only major disasters involving large rivers were considered [Mester et al., 2021, Bernhofen et al., 2018].
3. For a comprehensive global comparison study, different climate zones and continents should be considered [Dottori et al., 2016, Mester et al., 2021]. Therefore, floodplains from a set of four terrestrial biomes were chosen, exhibiting diverse ecological, climatological, topological, and morphological characteristics.
4. Flash flood events were excluded, as they would have failed the quality check of Tellman et al. [2021] due to the short timescales. Storm surge flooding, as well as floods resulting from mismanagement or failure of human-made structures, are not included; such flood types are beyond the scope of modeling capabilities for the GFM [Sauer et al., 2024, Risling et al., 2024].
5. All events had to fall within the timeframe of availability for the climate reanalysis product GSWP3-W5E5.

To select the regions, I examined the flood maps provided by the GFD and found 6 floodplains with repeated flooding events, in different regions and continents. Then I chose at maximum one event per year per region, always the largest flood event per year. The flood maps provided by Tellman et al. [2021] generally encompass a very large area, some even including the entire Indian subcontinent. Areas without relevant flooding are thus included in the maps. When choosing regions, I manually selected a subset of the original flood maps and determined

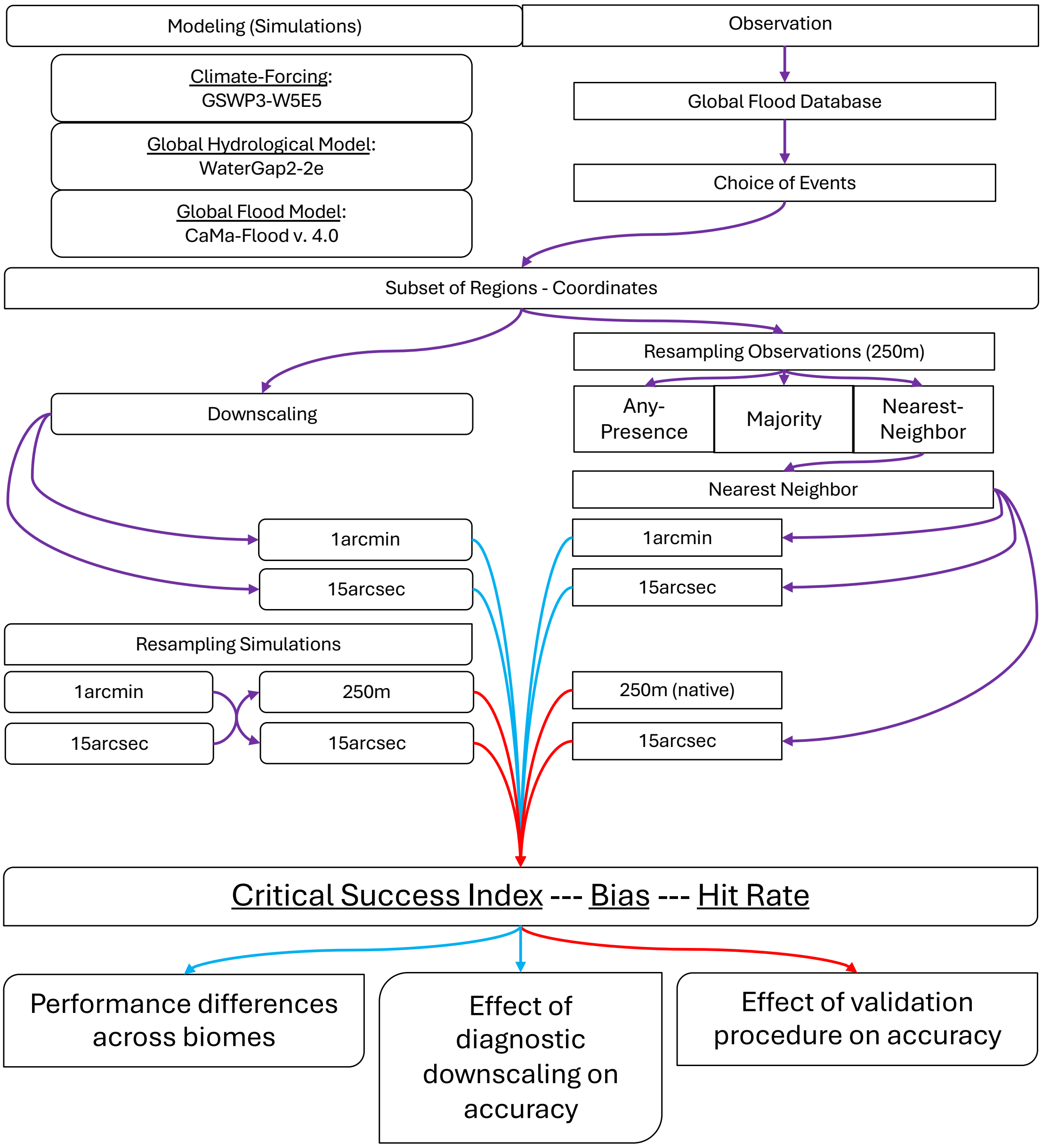

Modeling (Simulations)
Observation
Climate-Forcing: GSWP3-W5E5
Global Hydrological Model: WaterGap2-2e
Global Flood Model: CaMa-Flood v. 4.0
Global Flood Database
Choice of Events
Subset of Regions - Coordinates
Resampling Observations (250m)
Downscaling
Any-Presence
Majority
Nearest-Neighbor
Nearest Neighbor
1arcmin
15arcsec
1arcmin
15arcsec
Resampling Simulations
1arcmin
250m
15arcsec
15arcsec
250m (native)
15arcsec
Critical Success Index --- Bias --- Hit Rate
Performance differences across biomes
Effect of diagnostic downscaling on accuracy
Effect of validation procedure on accuracy

Figure 2: **An Overview of the Investigation Workflow:** The investigation workflow of observation and modeling, which meet in calculating performance scores. The analysis begins with a choice of events and subsequent generation of model flood discharge for these events. The resampling of observational data, sourced from the Global Flood Database, is tested with three resampling algorithms. Simulated flood extents are diagnostically downscaled to 1 arcmin and 15 arcsec, while observational data are resampled from 250 m using nearest-neighbor resampling. Simulated 1 arcmin is resampled to 15 arcsec, and simulated 15 arcsec is resampled to 250 m. Three analysis are executed based on three performance metrics. Events are further categorized into biomes.

the coordinates for each floodplain, focusing on the most relevant areas, relevance being characterized by the presence of flooding. Each event per floodplain received the same coordinates.

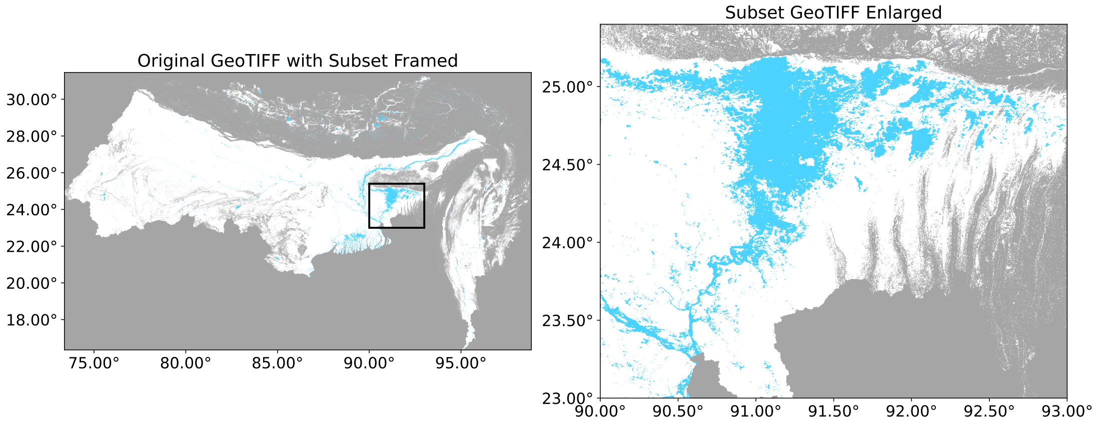


Figure 3: **Illustration of Need for a Region-Subset:** On the left, the full-scale map derived from the Global Flood Database is shown, with flood-affected areas rendered in light blue and regions with missing data represented in gray. The black-outlined frame defines the region selected for further analysis. On the right, an enlarged view of the framed subset is provided.

The regions chosen are the following: Bangladesh (3 Events), Assam, (3 Events), Northeast China (4 Events), Pakistan (2 Events), Bolivia (4 Events), and Myanmar (2 Events). A detailed description can be found in table 1 and the location is visualized in figures 4 and 5.

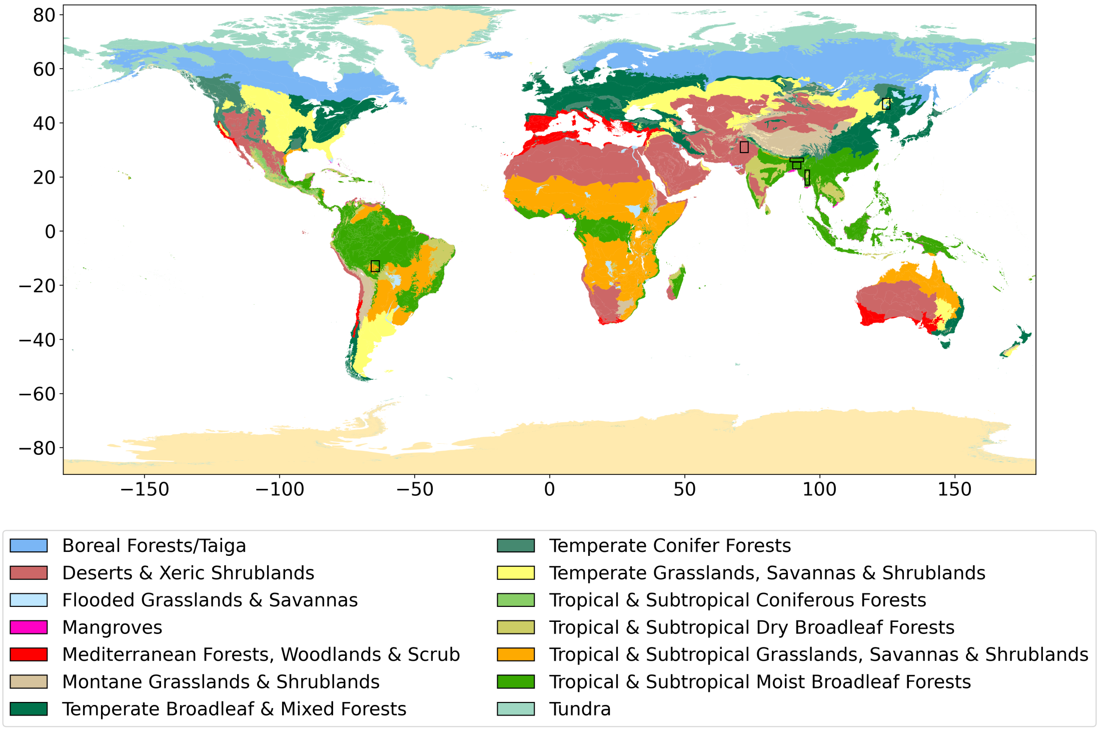
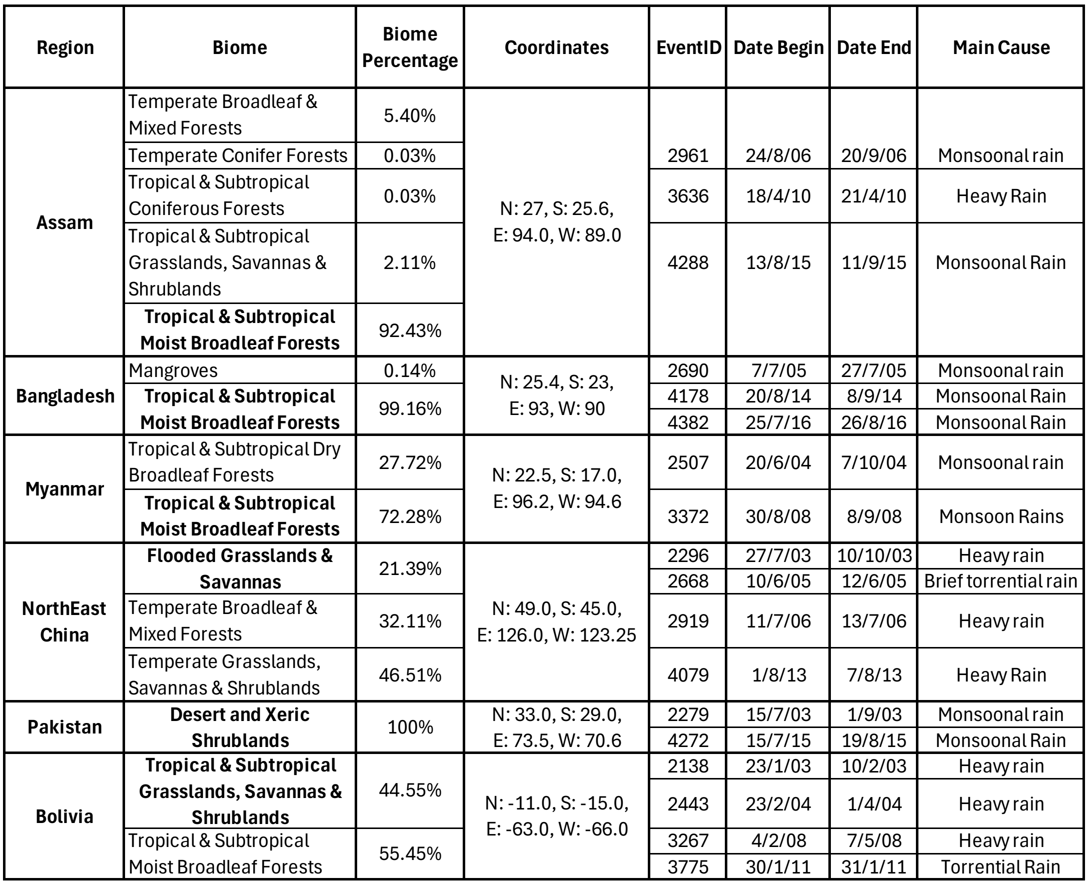

80
60
40
20
0
−20
−40
−60
−80
−150
−100
−50
0
50
100
150
Boreal Forests/Taiga
Deserts & Xeric Shrublands
Flooded Grasslands & Savannas
Mangroves
Mediterranean Forests, Woodlands & Scrub
Montane Grasslands & Shrublands
Temperate Broadleaf & Mixed Forests
Temperate Conifer Forests
Temperate Grasslands, Savannas & Shrublands
Tropical & Subtropical Coniferous Forests
Tropical & Subtropical Dry Broadleaf Forests
Tropical & Subtropical Grasslands, Savannas & Shrublands
Tropical & Subtropical Moist Broadleaf Forests
Tundra

Figure 4: **Global Biomes Map with Region Bounding Boxes:** Antarctica and Greenland are not mapped as biomes.

Four out of six selected floodplains are in South and South-East Asia. This



Region	Biome	Biome Percentage	Coordinates	EventID	Date Begin	Date End	Main Cause
Assam	Temperate Broadleaf & Mixed Forests	5.40%	N: 27, S: 25.6, E: 94.0, W: 89.0				
	Temperate Conifer Forests	0.03%		2961	24/8/06	20/9/06	Monsoonal rain
	Tropical & Subtropical Coniferous Forests	0.03%		3636	18/4/10	21/4/10	Heavy Rain
	Tropical & Subtropical Grasslands, Savannas & Shrublands	2.11%		4288	13/8/15	11/9/15	Monsoonal Rain
	Tropical & Subtropical Moist Broadleaf Forests	92.43%					
Bangladesh	Mangroves	0.14%	N: 25.4, S: 23, E: 93, W: 90	2690	7/7/05	27/7/05	Monsoonal rain
	Tropical & Subtropical Moist Broadleaf Forests	99.16%		4178	20/8/14	8/9/14	Monsoonal Rain
				4382	25/7/16	26/8/16	Monsoonal Rain
Myanmar	Tropical & Subtropical Dry Broadleaf Forests	27.72%	N: 22.5, S: 17.0, E: 96.2, W: 94.6	2507	20/6/04	7/10/04	Monsoonal rain
	Tropical & Subtropical Moist Broadleaf Forests	72.28%		3372	30/8/08	8/9/08	Monsoon Rains
NorthEast China	**Flooded Grasslands & Savannas**	21.39%	N: 49.0, S: 45.0, E: 126.0, W: 123.25	2296	27/7/03	10/10/03	Heavy rain
				2668	10/6/05	12/6/05	Brief torrential rain
	Temperate Broadleaf & Mixed Forests	32.11%		2919	11/7/06	13/7/06	Heavy rain
	Temperate Grasslands, Savannas & Shrublands	46.51%		4079	1/8/13	7/8/13	Heavy Rain
Pakistan	**Desert and Xeric Shrublands**	100%	N: 33.0, S: 29.0, E: 73.5, W: 70.6	2279	15/7/03	1/9/03	Monsoonal rain
				4272	15/7/15	19/8/15	Monsoonal Rain
Bolivia	**Tropical & Subtropical Grasslands, Savannas & Shrublands**	44.55%	N: -11.0, S: -15.0, E: -63.0, W: -66.0	2138	23/1/03	10/2/03	Heavy rain
				2443	23/2/04	1/4/04	Heavy rain
	Tropical & Subtropical Moist Broadleaf Forests	55.45%		3267	4/2/08	7/5/08	Heavy rain
				3775	30/1/11	31/1/11	Torrential Rain

Table 1: **Basic Information on All Flood Events.** EventIDs are taken from the Global Flood Database. Bolded rows in the Biome column denote the biome assigned to subsequent analysis.

is a matter of both availability and choice. Availability, because this is the region globally most prone to floods and exhibiting the highest exposure to floods [Tellman et al., 2021]. Choice for that reason specifically—these are the regions in which river flooding is most relevant.

The regions were classified into their respective biomes through a two-stage process. First, using the biome classification maps of Dinerstein et al. [2017], which are based on and extend the biome classifications of Olson et al. [2001], the presence of biomes within each delineated region was visually investigated. Second, the percentage of each biome within the region was calculated (table 1). Assigning a biome type to Assam, Bangladesh, and Pakistan was straightforward, as in each case one biome was dominant (> 90%). For Bolivia, Myanmar and Northeast China, the biomes were more evenly distributed. Here, I compared the biome maps with the flood maps, visually determined the areas where flooding was most prevalent, and then assigned the corresponding biome type. Four biomes emerged that will be investigated in this study: tropical and subtropical moist broadleaf forest (Assam, Myanmar, Bangladesh); desert and xeric shrublands (Pakistan);

tropical and subtropical grasslands, savannas and shrublands (Bolivia); flooded grasslands and savannas (Northeast China).

Tropical and subtropical moist broadleaf forests (tropical rainforests) are forests with closed canopies primarily consisting of broadleaf species [Dinerstein et al., 1995]. Deserts and xeric shrublands (deserts) exhibit sparse vegetation, the formation of soil crusts, and extensive impermeable surfaces [Lázaro et al., 2023]. Tropical and subtropical grasslands, savannas, and shrublands (tropical savannas) feature a varied physiognomy comprising both grasslands and woody vegetation [Hill et al., 2011]. Flooded grasslands and savannas (flooded grasslands) are regions frequently submerged underwater, characterized by a proliferation of grasses, periodic flooding, and a high water table for most of the year, which significantly affects vegetation [Joyce et al., 2016].

Resampling Methods

To compare observed floods with diagnostically downscaled simulations, GFD flood maps, obtained at 250 m resolution, were resampled to 15 arcsec and 1 arcmin resolution. This was done using the nearest-neighbor resampling method, as the GFD flood data are binary and nearest-neighbor resampling is suitable for categorical data [Sampson et al., 2015]. In nearest-neighbor resampling, each lower-resolution pixel receives either a flooded or non-flooded characterization by assigning it the value of the pixel in the finer-resolution grid that is nearest to its center. Majority resampling and any-presence resampling were also tested. In majority resampling, the values of the majority of the higher-resolution pixels within the lower-resolution pixel area are assigned to the lower-resolution pixel. In any-presence resampling, the lower-resolution flood cell is assigned flooded if any higher-resolution pixel within the lower-resolution pixel area is flooded. The validation plots of all three resampling techniques were inspected and the nearest-neighbor resampling was considered the most suitable (figure 6). Majority resampling provided maps comparable to those of nearest-neighbor resampling, but had a higher underestimation. In any-presence resampling, floods were noticeably overestimated compared to their original distribution.

In figure 6, the flood maps for any-presence resampling are corrupted. I tried several methods to resample using any-presence resampling (GDAL warp; Rasterio Resampling) and each time faced identical complications. Since the implications on flood distribution of resampling with this algorithm were nonetheless apparent and a continuation of work with this algorithm was thus out of the question, I neglected to fully fix the issue of visually mapping the respective flood data. Because of the relevance of the consequences of any-presence resampling to my

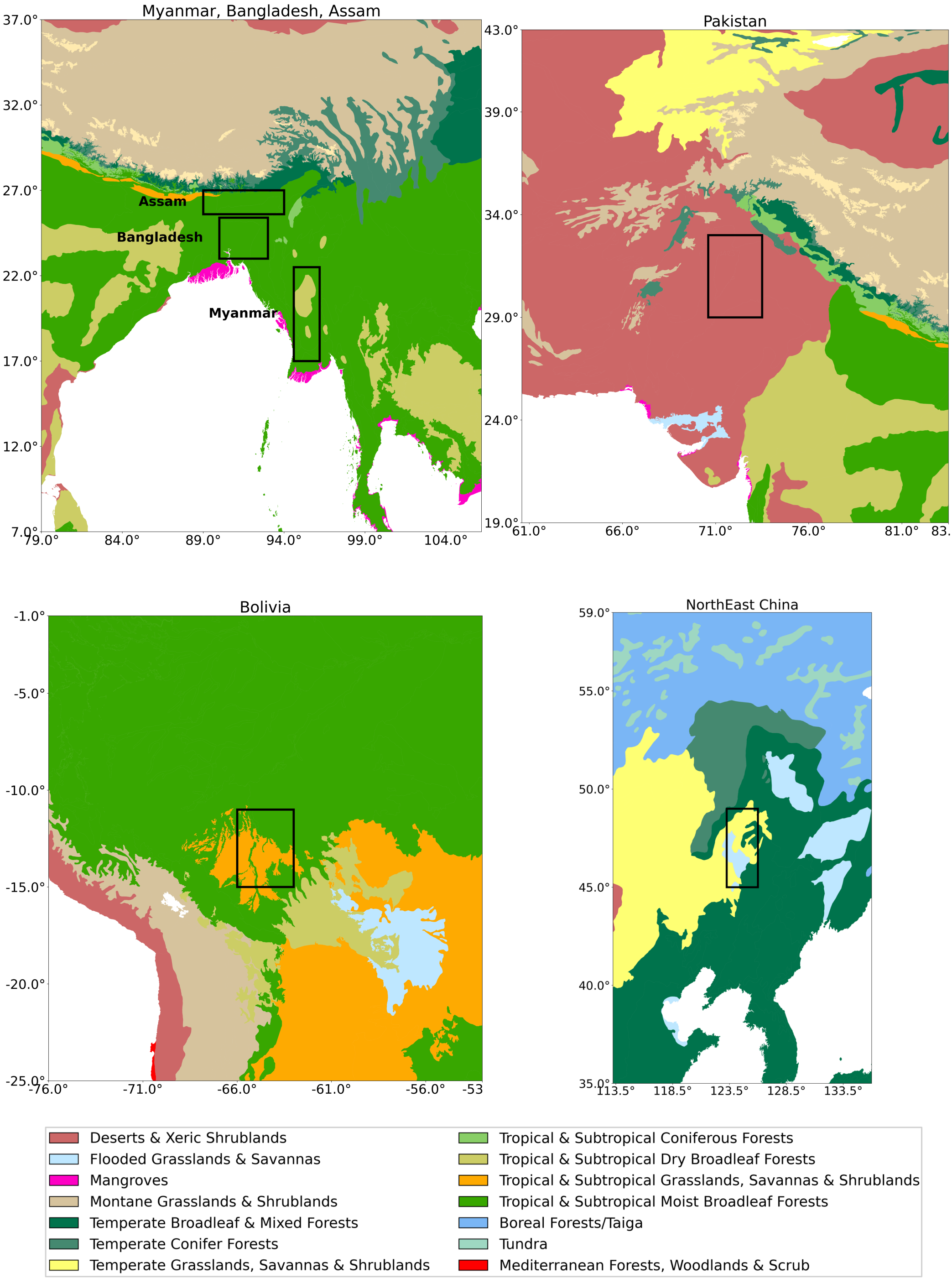

Myanmar, Bangladesh, Assam
37.0°
32.0°
27.0°
22.0°
17.0°
12.0°
7.0°
79.0°
84.0°
89.0°
94.0°
99.0°
104.0°
Assam
Bangladesh
Myanmar
Pakistan
43.0°
39.0°
34.0°
29.0°
24.0°
19.0°
61.0°
66.0°
71.0°
76.0°
81.0°
83.5°
Bolivia
-1.0°
-5.0°
-10.0°
-15.0°
-20.0°
-25.0°
-76.0°
-71.0°
-66.0°
-61.0°
-56.0°
-53
NorthEast China
59.0°
55.0°
50.0°
45.0°
40.0°
35.0°
113.5°
118.5°
123.5°
128.5°
133.5°
Deserts & Xeric Shrublands
Flooded Grasslands & Savannas
Mangroves
Montane Grasslands & Shrublands
Temperate Broadleaf & Mixed Forests
Temperate Conifer Forests
Temperate Grasslands, Savannas & Shrublands
Tropical & Subtropical Coniferous Forests
Tropical & Subtropical Dry Broadleaf Forests
Tropical & Subtropical Grasslands, Savannas & Shrublands
Tropical & Subtropical Moist Broadleaf Forests
Boreal Forests/Taiga
Tundra
Mediterranean Forests, Woodlands & Scrub

Figure 5: **Biome Maps of 6 Selected Regions:** The black-outlined frame defines the regions selected for further analysis.

choice of resampling algorithm, I nonetheless included the figures.

In general, resampling datasets has the potential to introduce errors related to false accuracy. However, since the datasets in question are binary, such errors are not introduced. This is because during resampling, interpolation between binary pixels does not produce new values, unlike when dealing with continuous datasets. Resampling could have introduced geospatial overlap errors, but these issues per-

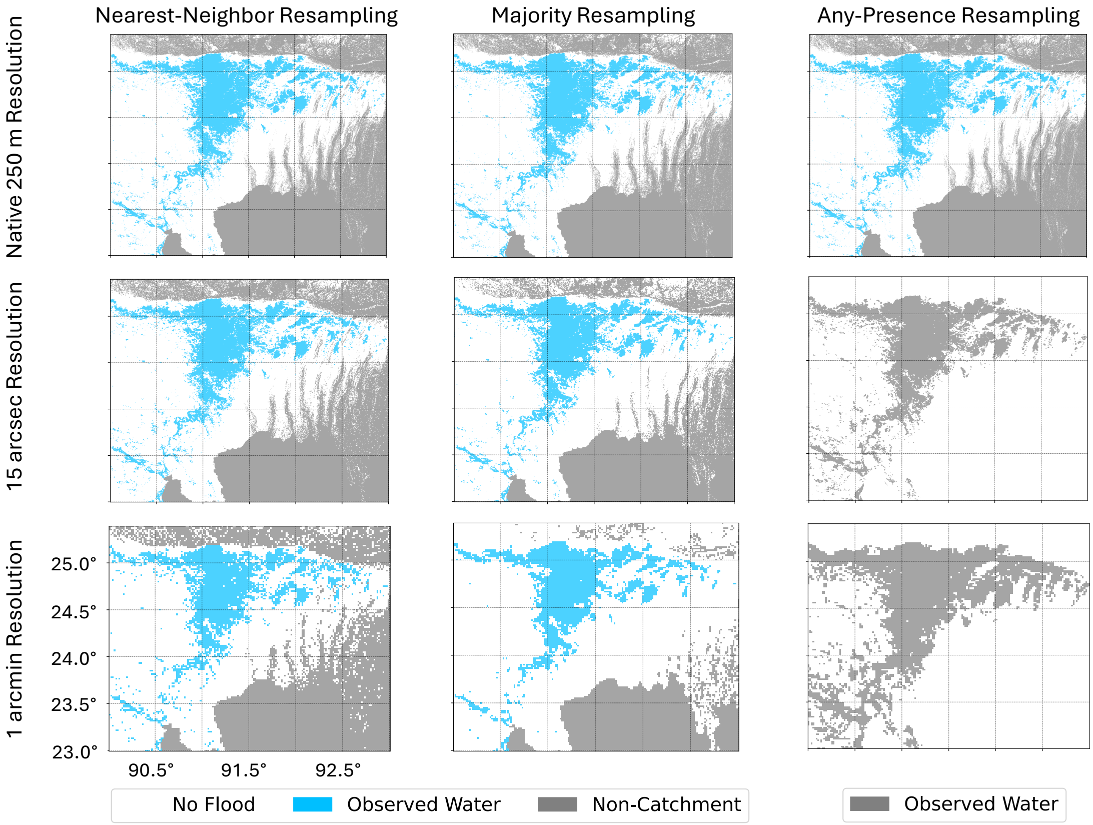


Figure 6: **Consequences of 3 Resampling Algorithms:** Flood maps for one flood event; at its native 250 m resolution (top row), resampled to 15 arcsec (middle row) and resampled to 1 arcmin (bottom row). Each column represents one resampling method. In the any-presence resampling flood maps of 1arcmin and 15arcsec, white space denotes no-flood and non-catchment, and gray areas represent observed water.

sist irrespective of the resolution to which the data is resampled. Nonetheless, they are unlikely to have influenced the validation outcomes [Bernhofen et al., 2018].

In conducting sensitivity analysis, which will be subsequently elaborated upon, I also resampled the simulated flood data from 1 arcmin to 15 arcsec, and from 15 arcsec to the native resolution of the GFD file, at 250 m. For both, I converted the continuous flood data from 1 arcmin and 15 arcsec respectively into binary flood data, using a flooded threshold of > 0. Then, I resampled the flood data using the nearest-neighbor method.

2.4 Analysis

For each region, event, and resolution (1 arcmin, 15 arcsec, 250 m), the flood extent simulated by CaMa-Flood will be evaluated against the flood extent of the GFD satellite observations [Tellman et al., 2021]. In addition, the results of these evaluations will be categorized into biomes.

To quantify model performance, the degree of overlap between the modeled flood extents and the observed DFO extents was calculated in terms of the number of pixels that were true positives, false positives, true negatives and false negatives. The CaMa-Flood outputs, originally expressed as extents with pixels representing flood depth, were transformed into binary water masks that indicate solely the extent of the flooding. Instead of a specific flood depth threshold, only the wet/dry threshold unique to each GFM output was used. The result is a binary grid where each cell is classified as flooded or not flooded for comparison purposes [Mester et al., 2021].

The numerical data from these calculations was then used to calculate three different performance scores—CSI, Bias and Hit Rate [Wilks, 2006]. These metrics are frequently employed in flood model evaluations and are used by several GFM providers for internal validation purposes [Bernhofen et al., 2018]. They were deemed suitable due to their exclusion of the dry area in the validation regions, which is beneficial in scenarios where correct ‘no’ forecasts are prevalent, such as in the extensive validation areas of this study [Bernhofen et al., 2018]. They were chosen as their results represent the most important aspects of model performance: model fit (CSI), model bias (Bias), and the proportion of total flood captured (Hit Rate) [Bernhofen et al., 2018]. The first score, and perceived as the most comprehensive [Bernhofen et al., 2018], is the critical success index (CSI) [Wilks, 2006]:

$$\text{CSI} = \frac{F_m \cap F_o}{F_m \cup F_o} \tag{2.1}$$

where Fm is the flooded area modeled by CaMa-Flood and Fo is the flooded area observed by satellite imagery. Fm ∩ Fo is the intersection area between the modeled and observed flood extent, i.e. the area correctly simulated as flooded by the model—true positives. Fm ∪ Fo is the union area of modeled and observed flooded extent—where either model or observation show flooding. CSI ranges from 0 to 1, where 1 represents a ’perfect fit’ of the model and penalizes overestimation [Sampson et al., 2015]. CSI is perceived as one of the most comprehensive scores [Bernhofen et al., 2018].

The second score is the Bias score [Wilks, 2006], which measures whether a forecast is biased towards underestimation or overestimation:

$$\text{Bias} = \frac{(F_m \cap F_o) + F_m}{(F_m \cap F_o) + F_o} - 1 \tag{2.2}$$

where Fm is the total modeled flood extent. A Bias score of 0 indicates an unbiased model. Positive and negative bias scores indicate bias towards overestimation and underestimation respectively. The Bias score rewards a large intersection area between modeled and observed flood extent.

The third score, the Hit Rate (HR) [Wilks, 2006], measures the proportion of the observed flood that was captured by the model:

$$\mathrm{HR} = \frac{F_m \cap F_o}{F_o} \tag{2.3}$$

Fm ∩ Fo again represent true positives and Fo is the total observed flood extent. The HR ranges from 1 (entire flood captured) to 0.

The analysis procedure concluded with a visual comparison of the simulated and the observed flood extent.

3. Results

In this section, I present the experimental findings of my three research objectives, beginning with the comparison of diagnostically downscaled simulations at 1 arcmin and 15 arcsec (primary analysis). I then compare diagnostic downscaling with nearest-neighbor resampling, aiming to distinguish the impact of spatial resolution from the method used to achieve it, and assess the influence of validation procedures (secondary analysis). And finally, I examine potential differences in performance across four biomes (tertiary analysis). Both quantitative metrics and qualitative observations are integrated to provide a comprehensive understanding of how resolution refinement influences model performance.

3.1 Impact of Diagnostic Downscaling

Figure 7 displays the distribution of performance scores derived from diagnostically downscaling the simulated flood data to resolutions of 1 arcmin and 15 arcsec.

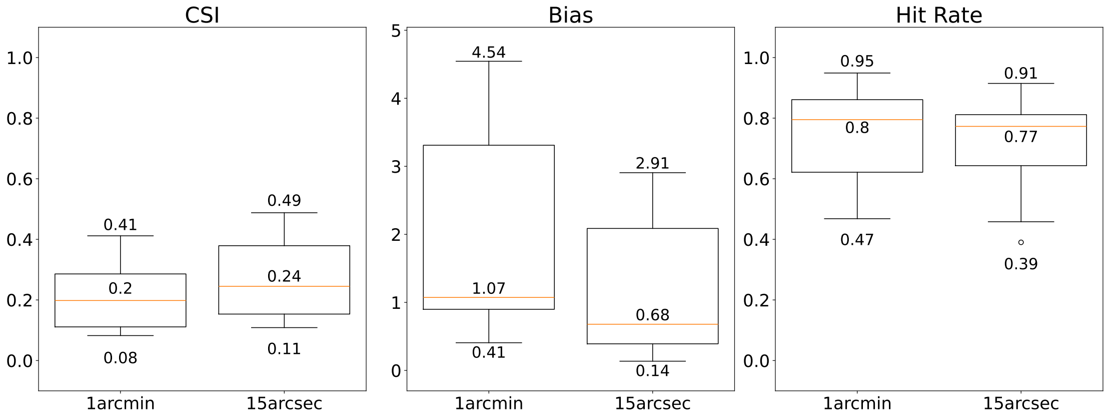


Figure 7: **Performance Metrics at 1 arcmin and 15 arcsec Resolution:** Boxplots display the distribution of performance scores—CSI (1 indicates perfect model fit), Bias (0 indicates no bias, positive values overestimation, negative values underestimation), and Hit Rate (1 indicates entire flood captured)—for 18 simulated flood events diagnostically downscaled to two spatial resolutions. Orange line, box, and whiskers represent median, interquartile range (IQR), and maximum or minium values, respectively; numbers indicate minimum, median, and maximum values.

The performances scores at 1 arcmin and 15 arcsec reveal substantial variation, both between resolutions and between flood events. Diagnostic downscaling to 15 arcsec produced higher CSI scores relative to 1 arcmin, with a top performance of 0.49 compared to 0.41 and a median of 0.24 relative to 0.2. This suggests that flood events were detected more accurately at finer resolution. Furthermore, Bias, which quantifies the tendency to underestimate flood extents (negative values) or overestimate (positive values) flood extents, was reduced at 15 arcsec compared to 1 arcmin, with a reduction in maximum Bias from 4.54 at 1 arcmin to 2.91 at 15 arcsec and a reduction in the median from 1.07 to 0.68. All flood events exhibited overestimation at both resolutions. Hit Rate scores show that between 47% and 95% of the floods are accurately predicted at 1 arcmin, as opposed to 39% to 91% at 15 arcsec, highlighting a reduction in the percentage of the observed flood accurately predicted. The reduction in false positive predictions indicated by the reduced Bias therefore also resulted in a decrease in true positives. Overall, however, the improvement in CSI scores indicates that the reduction in false positives outweighs the reduction in true positives.

Figure 8 presents a pivotal result of this study by illustrating the distribution of differences in performance metrics for each event between the 1 arcmin and 15 arcsec resolutions. For each flood event, the metric value obtained at 1 arcmin was subtracted from that at 15 arcsec. The analysis reveals that, for every event, CSI improved at 15 arcsec, while overestimation was reduced. Hit Rate was mostly reduced. Although the mean difference in each case in absolute terms is relatively low (CSI: 0.06, Bias: -0.74, Hit Rate: -0.03), a low standard deviation from the difference and consistent direction of differences result in highly statistically significant differences. The results of these statistical tests and those of the following comparisons are presented in table 2.

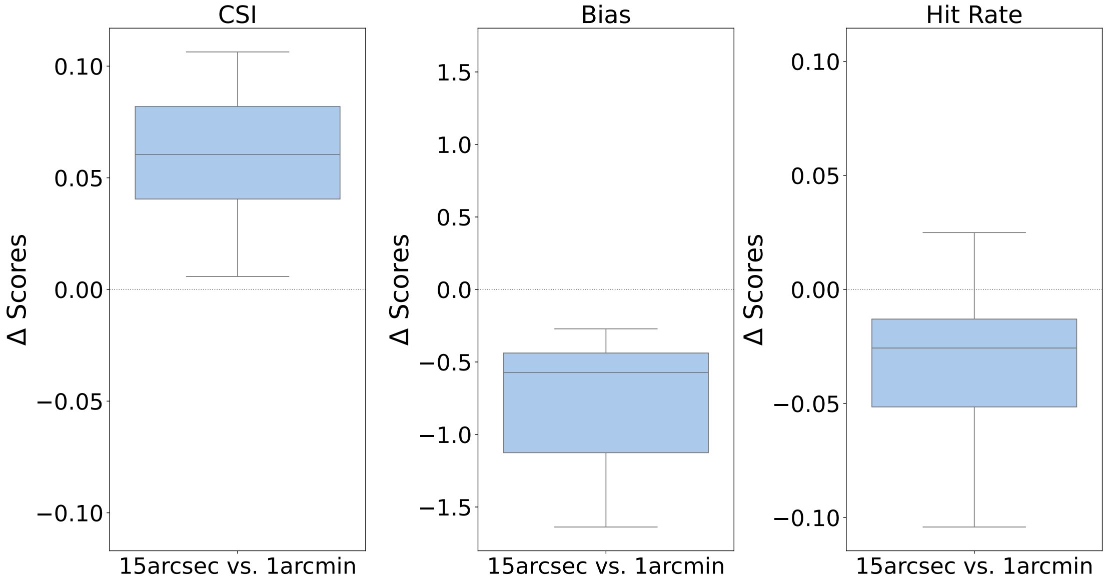


Figure 8: **Differences in Performance Metrics Between 1 arcmin and 15 arcsec Resolution:** Boxplots display the distribution of differences in performance metrics—CSI, Bias, and Hit Rate—between 18 simulated flood events diagnostically downscaled to 1 arcmin and 15 arcsec resolutions. For each flood event and metric, the difference (Δ Scores) was computed by subtracting the 1 arcmin score from the 15 arcsec score. The gray line at zero represents zero difference, positive values denote increase in scores, negative values a decrease. Gray line, box, and whiskers represent median, IQR, and maximum or minium values, respectively.

The decrease in overestimation is further illuminated by the qualitative comparisons shown in figure 9. The figure displays flood maps from three events, each representing a distinct biome, for both resolutions. Notably, the 1 arcmin maps exhibit a pronounced prevalence of dark blue areas, relative to 15 arcsec maps, which denote regions where water was simulated but not observed.

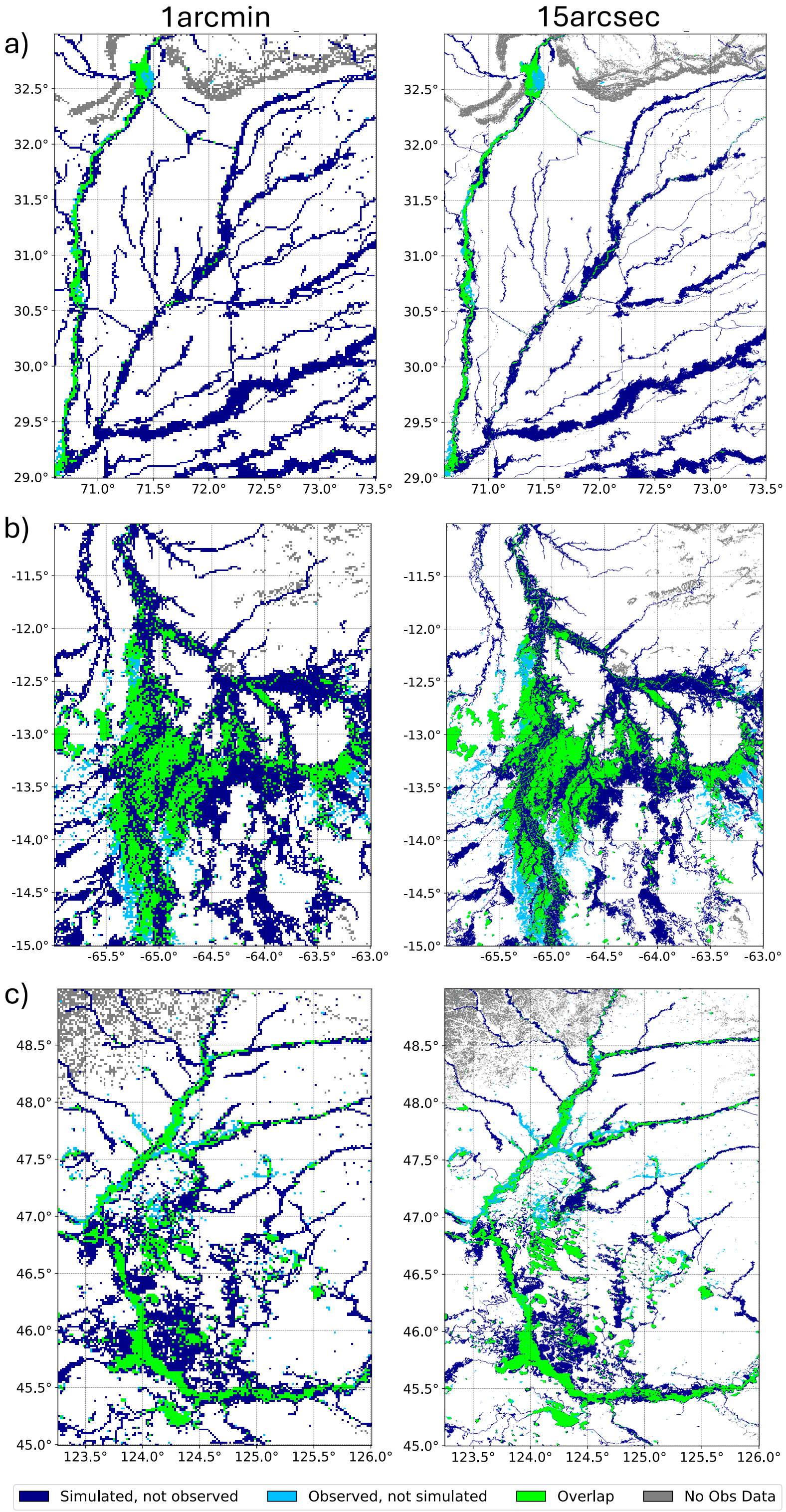


Figure 9: **Flood Maps for Three Flood Events at Resolutions of 1 arcmin and 15 arcsec:** Maps depict regions of concordance and discordance between simulated and observed water extents. **a)** Event ID 2279, Pakistan, (Biome: Desert); **b)** Event ID 3267, Bolivia (Biome: Tropical savanna); **c)** Event ID 2296, Northeast China (Biome: Flooded grasslands).

3.2 Impact of Resolution-change Methodology

Secondarily, I investigated whether the improved accuracy—specifically, the reduced overestimation observed at 15 arcsec—is a consequence of the diagnostic downscaling in CaMa-Flood or merely an artifact of computing scores at a finer spatial resolution. For this, I employed a two-pronged approach. First, I resampled the simulated flood data originally at 1 arcmin to 15 arcsec, instead of diagnostically downscaling them. Second, I resampled the simulated flood data at 15 arcsec to match the native 250 m resolution of the observational flood maps. In both cases, the continuous simulation outputs were first converted into binary flood maps and then resampled using the nearest-neighbor method.

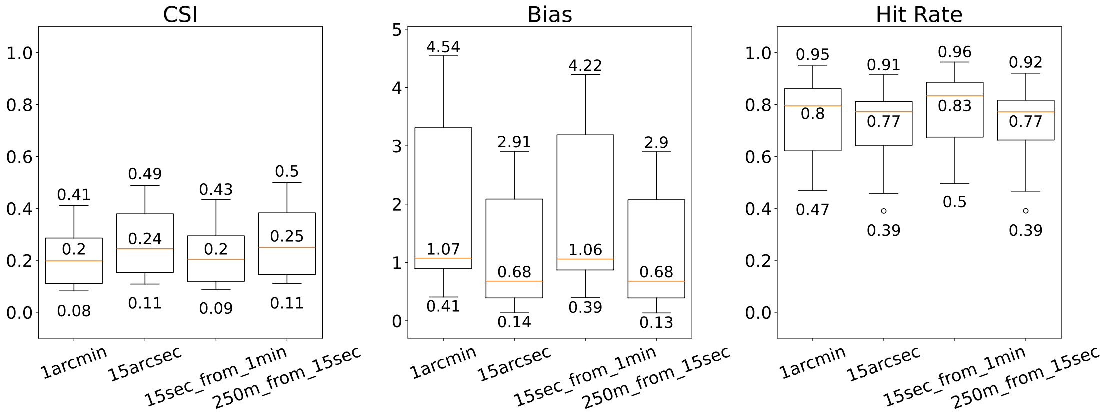


Figure 10: **Performance Metrics Across Three Resolutions and Two Resolution-change Methods:** Boxplots display the distribution of CSI (1 indicates perfect model fit), Bias (0 indicates no bias, positive values overestimation, negative values underestimation), and Hit Rate (1 indicates entire flood captured) scores for 18 simulated flood events across three resolutions and two resolution-change methods: Simulations diagnostically downscaled to 1 arcmin and 15 arcsec; simulation diagnostically downscaled to 1 arcmin and resampled to 15 arcsec (15sec_from_1min); simulation diagnostically downscaled to 15 arcsec and resampled to 250 m (250m_from_15sec). Orange line, box, and whiskers represent median, interquartile range (IQR), and minimum or maximum values, respectively; numbers indicate minimum, median, and maximum values. (Δ Scores)

Figures 10 and 11 present a second key result of this study. Figure 10 presents the distribution of scores for three resolutions (1 arcmin, 15 arcsec, 250 m) and two resolution-change methods (diagnostic downscaling, nearest-neighbor resampling). Figure 11 illustrates the differential effects of resolution and resolution-change method on performance metrics. For each analysis, the metric score from the second method is subtracted from that of the first method. CSI scores are enhanced and Bias scores decreased at 15 arcsec relative to 1 arcmin, even when the finer resolution is obtained via resampling (15_sec_from_1min). Otherwise stated: resampling flood data from 1 arcmin to 15 arcsec improves performance. The improvement from resampling to 15 arcsec is true for every flood event (boxplot B, figure 11). Resampling even increases Hit Rate, as opposed to diagnostic downscaling, where Hit Rate decreases at 15 arcsec. Moreover, these results are statistically significant (table 2). However, diagnostic downscaling outperforms resampling for every flood event in CSI and Bias (boxplot C). In fact, the difference between diagnostically downscaled flood data at 15 arcsec and resampled flood data at 15 arcsec, is only marginally lower than the difference between diagnostically downscaled data at 1 arcmin and 15 arcsec. Stated simply: resampled flood data at 15 arcsec is far closer to 1 arcmin accuracy than the diagnostically downscaled 15 arcsec accuracy.

Improvements are also observed when resampling from 15 arcsec to 250 m. A modest improvement in CSI at the 250 m resolution is observed, primarily driven by an increased Hit Rate, since the change in Bias hovers around 0. Combined with the improvement from resampling, this indicates that some improvement in performance is an artifact of computing scores at finer spatial resolutions.

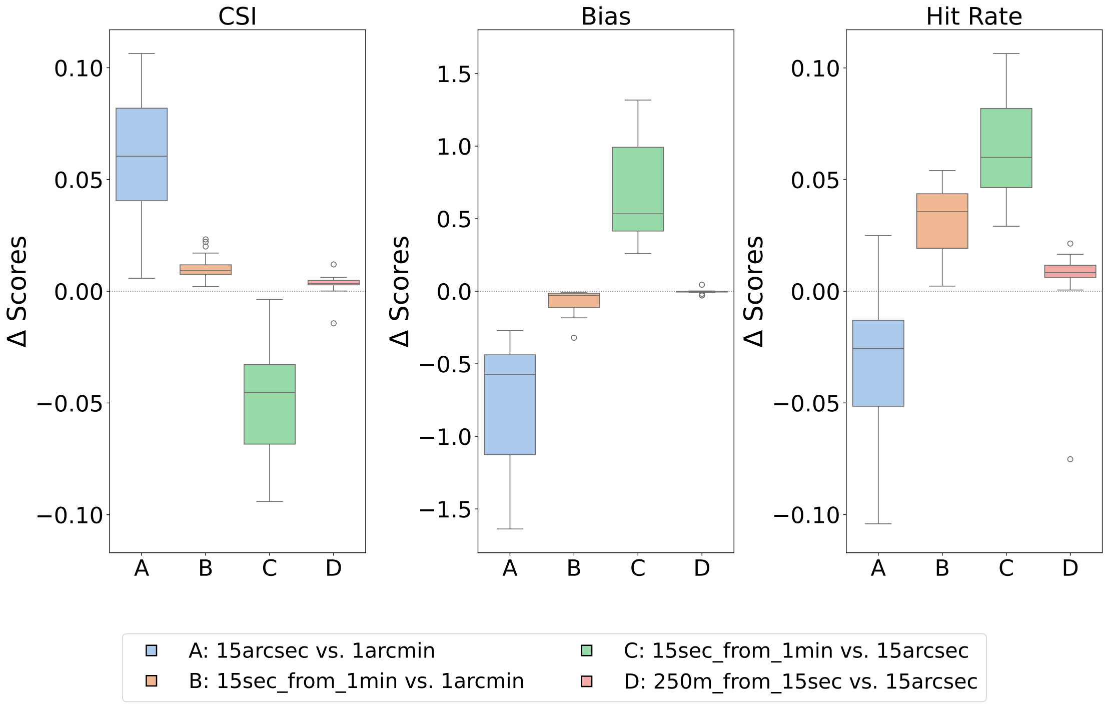


Figure 11: **Differences in Performance Metrics Across Three Resolutions and Two Resolution-change Methods:** Boxplots display the distribution of differences in performance metrics—CSI, Bias, and Hit Rate—between 18 simulated flood events across three resolutions and two resolution-change methods: Simulations diagnostically downscaled to 1 arcmin and 15 arcsec; simulation diagnostically downscaled to 1 arcmin and resampled to 15 arcsec (15sec_from_1min); simulation diagnostically downscaled to 15 arcsec and resampled to 250 m (250m_from_15sec). For each flood event and metric, the difference (Δ Scores) was computed by subtracting the scores of the second part of the versus equation in the legend from the first part. The gray line at zero represents zero difference, positive values denote increase in scores, negative values a decrease. gray line, box, and whiskers represent median, interquartile range (IQR), and maximum or minium values, respectively. The x-axis marks the comparison method.

Figure 12 indicates the mechanistic pathways of these results. When comparing simulated flood distributions, the disparity between 1 arcmin and 15 arcsec appears minimal if the change in resolution arises from resampling (**15secfrom1min**) rather than diagnostic downscaling (**15 arcsec**). There are, however, noticeable differences in the contours of the overlapping area: these show less pronounced edges of simulated and not observed areas at the resampled 15 arcsec (**15secfrom1min**) than 1 arcmin. A closer examination, cross-referenced with figures 3 and 6, indicates that the resampling of observational data, not the resampling of simulated data, is of relevance to performance. Resampling observational data to 1 arcmin leads to a noticeably more substantial deterioration in the fidelity of the observed flood delineation, than resampling only to 15 arcsec, particularly along the flood boundaries. This can be seen especially clearly in the gray areas of the flood map, which are part of the observed flood maps.

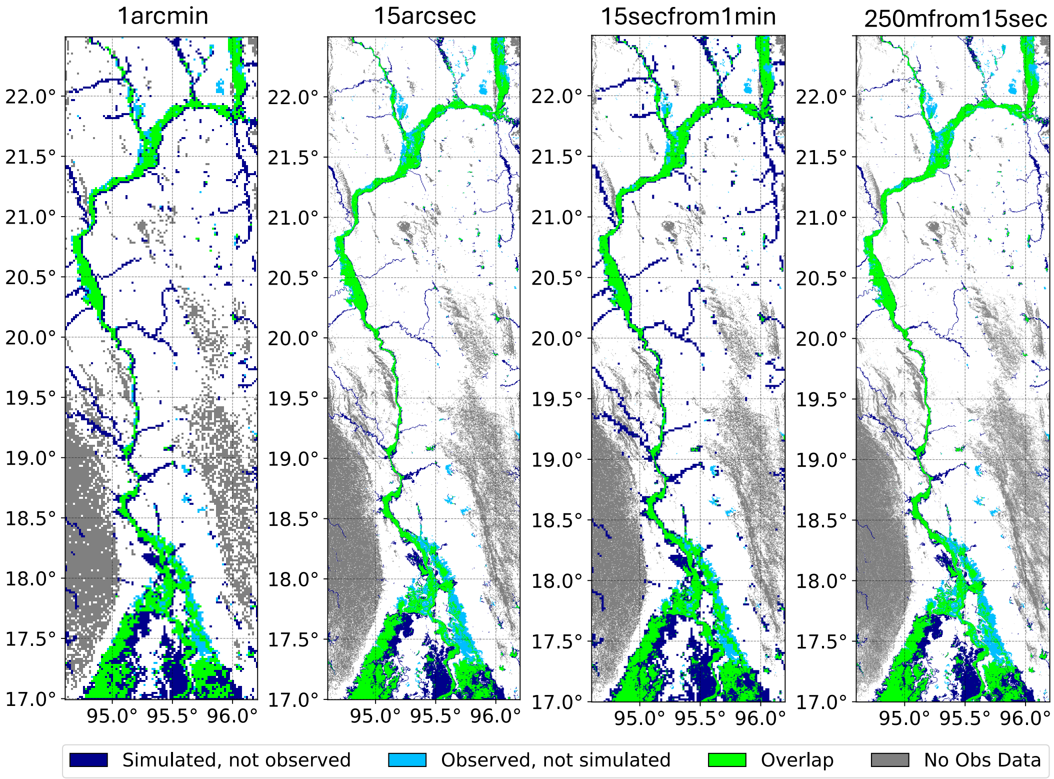
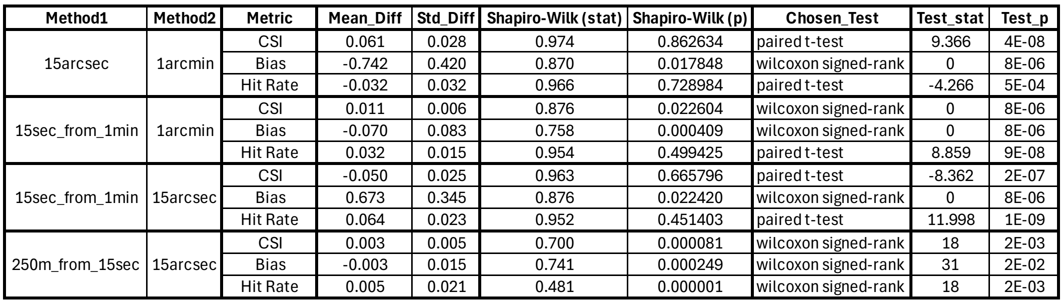


Figure 12: **Flood Maps of a Single Flood Event for Three Resolutions and Two Resolution-change Methods:** Maps depict regions of concordance and discordance between simulated and observed water extents. Resolutions and resolution-change methods are: simulations diagnostically downscaled to 1 arcmin and 15 arcsec; simulations diagnostically downscaled to 1 arcmin and resampled to 15 arcsec (15secfrom1min); simulations diagnostically downscaled to 15 arcsec and resampled to 250 m (250mfrom15sec). EventID: 2507, Myanmar, biome: Tropical rainforest.

Due to the small magnitude of the differences between the methods—on the order of hundredths or even thousandths—I evaluated whether these differences were statistically significant. Table 2 presents a comparative analysis of the methods described above across the three performance metrics.



Method1	Method2	Metric	Mean_Diff	Std_Diff	Shapiro-Wilk (stat)	Shapiro-Wilk (p)	Chosen_Test	Test_stat	Test_p
15arcsec	1arcmin	CSI	0.061	0.028	0.974	0.862634	paired t-test	9.366	4E-08
		Bias	-0.742	0.420	0.870	0.017848	wilcoxon signed-rank	0	8E-06
		Hit Rate	-0.032	0.032	0.966	0.728984	paired t-test	-4.266	5E-04
15sec_from_1min	1arcmin	CSI	0.011	0.006	0.876	0.022604	wilcoxon signed-rank	0	8E-06
		Bias	-0.070	0.083	0.758	0.000409	wilcoxon signed-rank	0	8E-06
		Hit Rate	0.032	0.015	0.954	0.499425	paired t-test	8.859	9E-08
15sec_from_1min	15arcsec	CSI	-0.050	0.025	0.963	0.665796	paired t-test	-8.362	2E-07
		Bias	0.673	0.345	0.876	0.022420	wilcoxon signed-rank	0	8E-06
		Hit Rate	0.064	0.023	0.952	0.451403	paired t-test	11.998	1E-09
250m_from_15sec	15arcsec	CSI	0.003	0.005	0.700	0.000081	wilcoxon signed-rank	18	2E-03
		Bias	-0.003	0.015	0.741	0.000249	wilcoxon signed-rank	31	2E-02
		Hit Rate	0.005	0.021	0.481	0.000001	wilcoxon signed-rank	18	2E-03

Table 2: **Statistical Tests Across Three Resolutions and Two Resolution-change Methods and Three Performance Metrics:** Mean_Diff represents the average difference between the events at the two methods, while Std_Diff denotes the corresponding standard deviation. Normality of the differences was evaluated using the Shapiro-Wilk test, with both the test statistic and p-value reported (Shapiro-Wilk (stat) and Shapiro-Wilk (p), respectively). Based on the normality results, either a paired t-test or the Wilcoxon Signed-Rank test was applied (Chosen_Test). Test_stat and Test_p indicate the test statistic and the p-value, respectively, for the selected test. For every comparison, n=18 events were compared.

Although many of the mean differences are on the order of hundredths or thousandths, the consistently low p-values indicate statistically significant differences across all comparisons. In particular, the 15 arcsec resolution outperforms the 1 arcmin resolution in terms of CSI (positive Mean_Diff) and shows a reduction in Bias (negative Mean_Diff). Additionally, 15 arcsec achieved through diagnostic downscaling yields higher CSI and lower Bias than 15 arcsec derived via resampling. While the absolute differences are small, the tight clustering of these differences (as reflected in low Std_Diff values) contributes to high test statistics and correspondingly low p-values.

3.3 Performance Across Biomes

Finally, addressing my tertiary research question, I investigated whether the benefits of downscaling vary across different ecological contexts as defined by biomes.

I first calculated the mean across all events for each performance metric, and then computed the distance to this mean for each event and categorized this into biomes. The results of this can be seen in figure 13 for 1 arcmin and 15 arcsec.

At both resolutions, the Critical Success Index (CSI) tends to be higher than the overall mean in the tropical rainforest biome, with a median score of 0.1 (1 arcmin) and 0.12 (15 arcsec) above the overall mean. Deserts and tropical savannas on average perform noticeably below the mean. Flooded grasslands also fall below the mean in most events, although with a lesser deviation. Similarly, Bias (overestimation) is consistently below the mean in the tropical rainforest and flooded grasslands biome and above the mean in the other two biomes. This holds for both resolutions. The Hit Rate exhibits a slightly different pattern: here desert and tropical rainforest events remain above the mean, whereas flooded grasslands and tropical savannas fall below the average Hit Rate. However, it should be noted that the number of events differs considerably across biomes—8 in tropical rainforest versus only 4 in flooded grasslands and tropical savannas and 2 in the desert biome. This may influence the reliability of these comparisons.

Subsequently, I evaluated how performance metrics across these biomes are affected by diagnostic downscaling. Figure 2 presents the differences between the simulated flood data downscaled to 1 arcmin and 15 arcsec, stratified by biome.

First off, performance for every event in every biome (as calculated in CSI and Bias) improves at 15arcsec relative to 1arcmin. This is a restatement of the findings from section 3.1. Noticeably, the median increase in CSI when diagnostically downscaling is highest in the tropical rainforest biome, whereas desert and tropical savannas experience the least improvement. Flooded grasslands are again located between tropical rainforests and the other two biomes. The increase in CSI is highest where the absolute scores were highest. For Bias, on the other hand, the decrease is most marked in deserts and tropical savannas, between -1 and -1.5, while overestimation (Bias) only decreases by a median of about 0.5 in tropical rainforests and flooded grasslands. Furthermore, the Hit Rate declines most markedly in the tropical rainforest biome compared to the desert and flooded grasslands, while in the tropical savannas the Hit Rate even increases. This initially perplexing discrepancy—less reduction in overestimation and a decrease in Hit Rate, yet the most substantial increase in CSI for tropical rainforest biome follows from the logic through which these scores are calculated. This logic and its mechanistic implications will be examined in the ensuing discussion section.

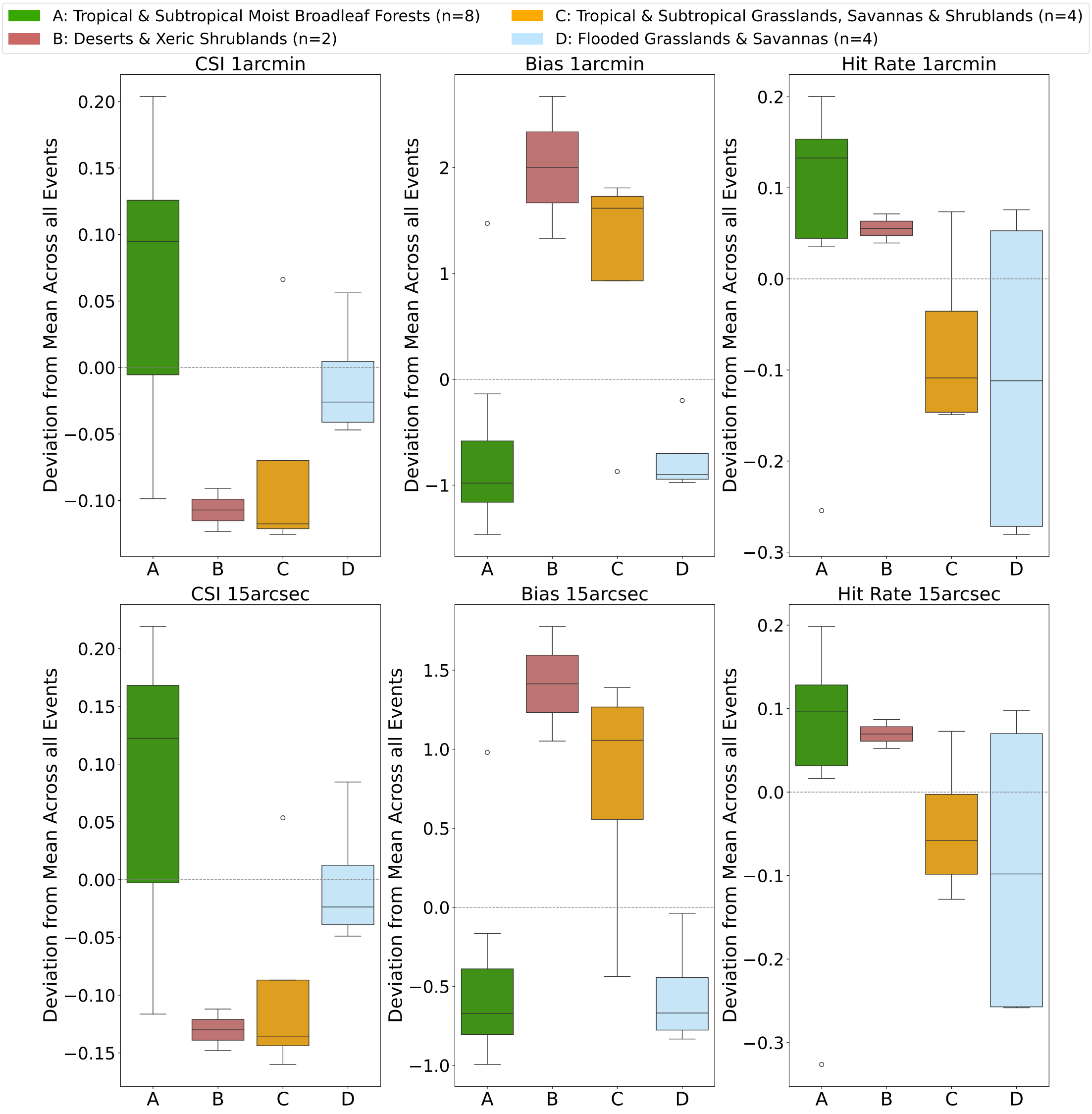


Figure 13: **Performance Metrics at Two Resolutions Stratified by Biome:** Boxplots display the distribution of scores—CSI (1 indicates perfect model fit), Bias (0 indicates no bias, positive values overestimation, negative values underestimation), and Hit Rate (1 indicates entire flood captured)—between 18 simulated flood events. Simulations diagnostically downscaled to 1 arcmin (top row) and 15 arcsec (bottom row). Each event's deviation from the overall mean across all biomes is shown, stratified into biomes. Each panel represents one performance metric. The horizontal zero line marks no deviation from the overall mean score, positive values denote above average scores, negative values below average scores. gray line, box, and whiskers represent median, interquartile range (IQR), and maximum or minium values, respectively.

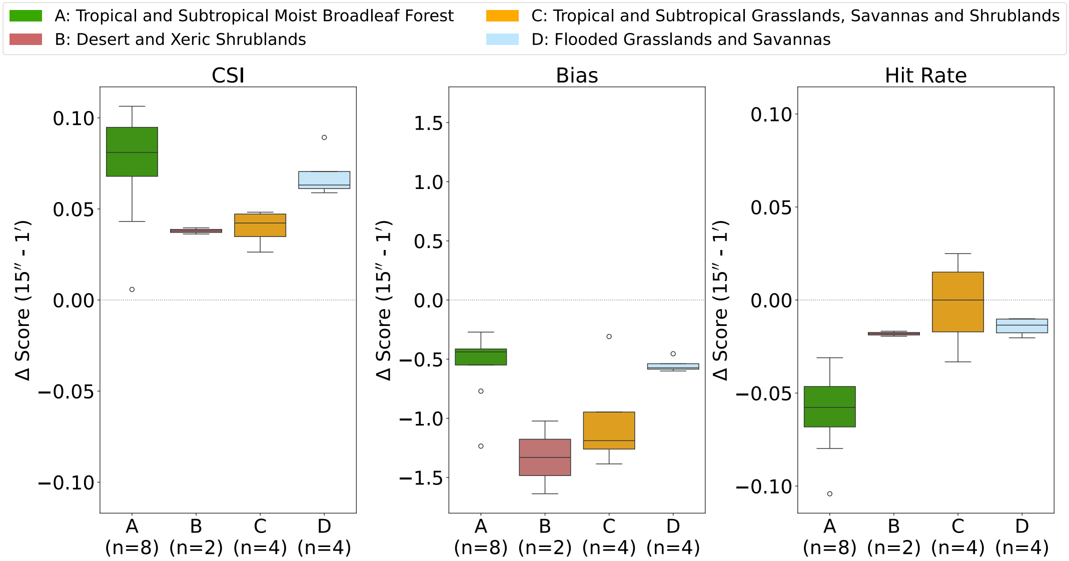


Figure 14: **Differences in Performance Metrics Between 1 arcmin and 15 arcsec Stratified by Biome:** Boxplots display the distribution of differences in performance metrics—CSI(1 indicates perfect model fit), Bias (0 indicates no bias, positive values overestimation, negative values underestimation),and Hit Rate (1 indicates entire flood captured)—between 18 simulated flood events diagnostically downscaled to 1 arcmin and 15 arcsec resolutions. For each flood event and metric, the difference (Δ Scores) was computed by subtracting the 1 arcmin score from the 15 arcsec score and subsequently stratified into biomes. The gray line at zero represents zero difference, positive values denote increase in scores, negative values a decrease. gray line, box, and whiskers represent median, interquartile range (IQR), and maximum or minium values, respectively.

4. Discussion

This study primarily evaluated the efficacy of diagnostic downscaling techniques within the CaMa-Flood global flood model, while also elucidating the differential impact of diagnostic downscaling versus resampling and finally exploring the potential variability in performance across diverse biomes. In addressing these objectives, the research illuminated several key findings, raised important methodological considerations, and pointed toward new directions for future research, reinforcing the need for continued refinement in flood modeling approaches and cross-disciplinary investigation.

4.1 Key Results

In summary, three key results emerged:

1. **Diagnostic Downscaling Improves Performance**: Downscaling simulated flood data to a 15 arcsec resolution markedly enhances model accuracy compared to using a 1 arcmin resolution for every flood event. The 15 arcsec approach delivers higher CSI scores and notably reduces overestimation, although it also shows a slight decrease in Hit Rate.

2. **Validation Procedures Matter**: Resampling to a higher resolution also improves model accuracy. However, this results not from a gain through resampling, but rather from a reduction in resolution-induced loss of spatial fidelity of the observational data through resampling from its native 250 m resolution to 15 arcsec relative to 1 arcmin. Furthermore, diagnostic downscaling consistently outperforms resampling.

3. **Performance Differs Across Biomes**: When stratified by biome, the overall performance and benefits of downscaling are most pronounced in the tropical rainforest biome, followed by flooded grasslands. In contrast, deserts and tropical savannas show relatively lower performance and performance improvements.

The purpose of this discussion is to interpret and contextualize the experimental findings by critically evaluating how diagnostic downscaling at different resolutions affects flood model performance. This section integrates both quantitative metrics and qualitative observations, situating the results within the broader literature on flood validation and resolution refinement. It will focus on examining the nuances introduced through validation procedures, and attempt to define potential mechanistic pathways to explain the results. In addition, the discussion will explore the differential impacts observed in different biomes, thereby assessing the methodological implications for flood modeling across varied ecological contexts. By addressing these aspects, I seek to elucidate the strengths and limitations of the study while discussing the potential for future methodological enhancements in the field. Due to the interconnected nature of the primary and secondary analysis, I will begin by discussing these in concert and then separately discuss the results of the tertiary analysis.

4.2 Diagnostic Downscaling Improves Performance and Validation Procedures Matter

Diagnostic Downscaling Improves Performance

The observed improvement in CSI at 15 arcsec resolution (figures 7 and 8) suggests that finer resolutions enable a more accurate delineation of flood extents. Equally notable is the observed reduction in Bias at 15 arcsec, which indicates that the overestimation of flooded areas is diminished when higher spatial detail is incorporated (figure 9). This is likely due to the model's ability to better capture topographic variations and local hydrodynamic processes on a sub-grid scale [Yamazaki et al., 2011, 2019]. This also makes intuitive sense. If at 1 arcmin a pixel is marked as flooded, at 15 arcsec there are 4 possible pixels to which to assign values, potentially reducing the flooded pixels by 3/4. A slight decrease in the Hit Rate (true positives) was observed, which appears to be a trade-off for the decrease in false positives provided diagnostic downscaling. Given that CSI is considered a comprehensive measure of model performance, in general these results highlight the potential advantages of employing a finer resolution for diagnostic downscaling in flood modeling [Bernhofen et al., 2018].

More specifically, although false positives are notably minimized at 15 arcsec (Bias), this correlates with an increase in false negatives (Hit Rate). The cost-benefit ratio of further diagnostically downscaling CaMa-Flood can thus be stated as a cost of the amplification of false negatives at the benefit of a reduction in false positives. The significant improvement in CSI, taken to be the most comprehensive score, would indicate this as a worthwhile trade-off. Nevertheless, the nature of the trade-off necessitates a deeper questioning of its practicality because, in terms of practical decision making, the cost of a false negative can be orders of magnitude higher than the cost of a false positive. When individuals are convinced of their safety, they are likely to invest in infrastructure and maintain their lives in specific regions. If this confidence is misplaced, the consequences can be catastrophic and irreversible. Conversely, the perception of insecurity in an area will deter people from developing infrastructure and settling there. Although this decision may involve expenses related to relocating or establishing a livelihood in less ideal circumstances, these are marginal compared to the devastating outcomes of incorrectly assuming safety.

This consideration needs to be taken into serious account when deciding how to interpret the results of this study. The pivotal question is: what are the intended applications of global flood models? Is the goal to eventually make locally calibrated models unnecessary, offering models that are cost-effective, user-friendly, and uniformly reliable? Achieving this would necessitate higher resolutions, but the trade-off highlighted in this study must be thoroughly evaluated and addressed. Conversely, if GFMs are designed for large-scale evaluations, planning, and research, where aggregate data is prioritized over specific details, higher resolutions already offer distinct and statistically significant advantages.

These findings are consistent with the latest advances in flood modeling that underscore the importance of high-resolution data and the crucial role of spatial resolution in the level of agreement between GFM and observational data [Bryant et al., 2024, Wing et al., 2024, Risling et al., 2024, Sampson et al., 2015]. The literature indicates that more detailed spatial resolutions can more accurately capture flood extents, emphasizing the critical role of accurate DEMs in enhancing model performance. For example, a study by Fereshtehpour et al. [2024] found that coarser DEMs led to a decreased accuracy in flood depth predictions and that even a slight improvement in data resolution in data-scarce regions could provide significant added value, ultimately improving flood risk management.

It should be noted that, generally, the CSI scores in this study are on the lower side, indicating poor overall agreement. Previous GFM validation studies have shown that in specific areas certain GFMs can achieve CSI scores as high as 0.9, although they can fall to as low as 0.02 [Bernhofen et al., 2018, Risling et al., 2024, Mester et al., 2021]. Values greater than 0.7 are considered good, while those less than 0.5 are considered poor [Bernhofen et al., 2018]. Low CSI scores can be traced to high Bias scores—denoting overestimation—across all events and resolutions. This overestimation can be attributed to several factors. First, GHM are known to overestimate river discharge [Heinicke et al., 2024]. Second, the flood discharge from CaMa-Flood was calculated by assigning to each high-resolution grid cell the annual maximum daily value, without any flood protection measures [Mester et al., 2021]. Although only the largest flood event per year was chosen from the GFD for each respective region, this method of attaining simulated flood values still leads to overestimation. This study thus substantiates the already well-documented need and difficulty of incorporating flood mitigation measures, highlighting the constraints of GFMs where such measures have been established [Risling et al., 2024]. Third, I implemented a straightforward threshold to assign a flooded value to the binary simulated flood mask: Any flood depth greater than 0 was marked as flooded. Fourth, the GFD is based on MODIS imagery, which encounters challenges such as cloud cover, canopy interference, and terrain shadows—issues that are particularly problematic in areas with low sun angles or uneven landscapes. Although the GFD omits regions with slopes greater than 5 ° and uses multi-day compositing, this multi-day compositing along with possible cloud coverage, might result in an underestimation of flood extent [Tellman et al., 2021, Sauer et al., 2024]. Finally, when water channels are narrower than than MODIS's 250 m resolution, they could be simulated yet not observed, leading to overestimation.

Validation Procedures Matter

I found that resampling the simulated data to a higher resolution with the nearest-neighbor resampling method consistently increases performance in all flood events (figure 10). However, although both resampling and diagnostic downscaling enhance model outputs at 15 arcsec compared to the 1 arcmin resolution, diagnostic downscaling consistently produces higher CSI values and a more pronounced reduction in Bias (figure 11). This result is indicative of the mechanistic advantages offered by diagnostic methods; by explicitly redistributing flood volume according to the underlying elevation data, these methods preserve the inherent hydrological signals that are absent in resampling approaches. In both scenarios, a single 1 arcmin pixel is subdivided into four 15 arcsec pixels. However, while diagnostic downscaling determines the flooded status of each smaller pixel based on detailed information from the digital elevation model, resampling simply assigns values based on the nearest larger pixel on the lower resolution grid. In cases of perfect grid alignment, all four smaller pixels will inherit the value of the original larger pixel (figure 12).

Therefore, the simulated flood map at 15 arcsec when resampled from 1 arcmin is identical to its original state (figure 12). However, the performance scores differ (figure 11). A possible explanation for this phenomenon can be identified through a qualitative analysis of flood maps. The discrepancy in scores arises not because of differences in the simulated flood map, but in the observed flood map between 15 arcsec and 1 arcmin. At 15 arcsec, the observed data are reduced only by a factor of approximately two relative to its native resolution of 250 m, whereas at 1 arcmin the resolution is reduced by approximately a factor of six. Consequently, the 15 arcsec resolution exhibits less degradation of the native flood map fidelity.

Resampling can thus compromise data fidelity; as the disparity between native resolution and final resolution increases, the data increasingly deviates from their original form (figures 6 and 12). It is therefore important to add that the differences in the scores for the diagnostically downscaled resolutions cannot be solely attributed to the performance of the downscaling. The impact of resampling the observed data must also be considered. However, the remaining difference between diagnostically downscaling to 15 arcsec vs. resampling to 15 arcsec supports the evidence that diagnostic downscaling still improves the actual accuracy significantly.

In addition, these results support the idea that care should be taken when labeling remotely detected flood maps as ‘ground truth’. The challenges outlined here concerning resampling are just one part of the complexity. Remote sensing of floods is also inherently limited by various factors, including unique issues posed by different satellite sensors, the spatial resolution of the collected data, the timing of these acquisitions, and the assumptions and constraints involved in the subsequent analysis of change detection [Risling et al., 2024].

Statistical Significance

Although the mean differences between all validation procedures are relatively small, on the order of a hundredth or a thousandth, the p-values are exceptionally low, denoting statistically significant differences between all metrics and methods (table 2). There are two reasons for this: first, the differences consistently favor one method over another—nearly every event shows a small but directional difference. Moreover, the low standard deviations observed in most comparisons show that the differences are consistent and tightly clustered, thus enhancing the significance of these mean differences. Under such conditions, statistical tests, such as the paired t-test or the Wilcoxon signed-rank test, reliably detect a deviation from zero. Thus, even though the absolute differences are small, the high consistency across events results in high test statistics and correspondingly low p-values, underscoring a robust statistical distinction between the methods.

However, it is important to note that statistical significance does not necessarily translate into practical significance. Although the differences may be highly unlikely to have occurred by chance, their real-world impact may be negligible, since the effect size is very small. In a decision-making context, what ultimately matters is whether the difference is large enough to influence outcomes, resource allocation, or policy [Ward et al., 2015]. Thus, even though the differences in resolution when achieved through resampling are highly statistically significant, my discussion on the origin of the difference and the small effect size support the claim that resampling does not actually result in enhanced accuracy. Diagnostic downscaling, on the other hand, produces a larger effect size, traceable to mechanistic advantages. This potentially translates into practical relevance, though requiring the aforementioned considerations concerning the trade-off between false positives and false negatives.

Limitations

Although the present study makes strides in evaluating diagnostic downscaling methods and validation procedures for global flood modeling, several methodological and practical limitations warrant careful consideration.

The focus of my study on a single combination of climate forcing and global hydrological model to force CaMa-Flood limits the ability to generalize the results. The diagnostic downscaling process is executed on the CaMa-Flood discharge and not the GHM input, so it is not immediately apparent how a difference in the effect of downscaling would emerge. However, Mester et al. [2021] showed that the combination of GHM and climate forcing can lead to variation in the estimate of flood depth and inundation areas, with different combinations of GHM and climate forcing favoring distinct areas. If these regions are consistently topographically diverse, downscaling might have varying effects, potentially improving the precision under certain GHM-climate forcing combinations more effectively than others. Therefore, future work could benefit from incorporating multiple datasets and models to assess the robustness of diagnostic downscaling under varying hydrometeorological conditions. Furthermore, it was beyond the scope of this study to incorporate flood protection measures into the modeling process and analyze how these would impact diagnostic downscaling. This might influence the effect of diagnostic downscaling in a comparable manner to using various combinations of GHM-climate forcing, through alterations in inundation areas. Future research should explicitly address this issue by incorporating flood protection measures into the modeling process.

When comparing simulated and observed flood extents, the flood depth was not taken into account. The simulated flood data only register the occurrence of flooding—any value above 0 is treated the same—without distinguishing between shallow flooding (e.g., 1 centimeter) and much deeper flooding (e.g., 10 meters). In practical terms, this is an influential difference. Converting continuous simulation outputs into binary maps, or maintaining binary observed flood maps, thus conceals valuable information. This shortcoming might influence two parts of this study: the extent to which resampling is inferior to diagnostic downscaling and the extent to which the observed data fidelity is reduced at coarser resolutions.

Diagnostic downscaling to 15 arcsec considers flood depth, whereas the resampling technique used here (conversion of continuous flood depth data to a binary flood mask, which was subsequently resampled using nearest-neighbor) does not. Maintaining simulated flood data with continuous flood depth values and using continuous resampling techniques could more closely approximate diagnostic downscaling. In continuous resampling methods, such as bilinear or cubic interpolation, the value assigned to each resampled pixel is calculated as a weighted average of nearby pixels. However, it is important to note that no resampling method will transform a small non-zero value (e.g., 0.00001) into zero; if even one contributing pixel has a small non-zero value (like 0.00001), the resulting average will also be non-zero rather than being set to zero outright. Therefore, to potentially minimize the distance between diagnostic downscaling and resampling, continuous resampling techniques would need to be used in conjunction with a revised flood threshold (for instance, setting it at a more practical level like 0.15 instead of 0). Then, for example, bilinear resampling might reduce a value from 0.32 to 0.08, potentially excluding it from being classified as flooded. This hypothesis requires testing, and even if relatively more successful than nearest-neighbor resampling, I assume the overall performance is unlikely to be improved to the same degree as with diagnostic downscaling.

More pronounced benefits to validation could be seen if continuous resampling methods could be used on an observational dataset with continuous flood-depth data. Then, potentially, continuous resampling to lower resolutions could reduce errors and provide a more nuanced representation of the underlying hydrodynamic processes than the nearest-neighbor approach. In other words, the fidelity of flood maps at 1 arcmin could potentially be less impaired than currently is the case.

Finally, this study did not investigate the highest resolution achievable by diagnostic downscaling in CaMa-Flood, 3 arcsec (90 m). This definitely requires continued research in the future. However, the results of this study have shown that a comparison to the 3 arcsec simulation using the 250 m observational dataset is not straightforward and might not produce reliable results, since the 250 m raster would have required resampling, inherent to which would be the limitations discussed here. Therefore, future research should search for a comparison dataset that is close in native accuracy to the 90 m resolution. This in turn would not address the issue of comparing the performance of the 3 arcsec diagnostic downscaling to 15 arcsec, or 1 arcmin diagnostic downscaling, as the observational data would again require resampling. Using separate observational datasets for comparison is a possible solution to this problem. However, that would introduce its own uncertainties, as distinct observational datasets are subject to separate limitations and uncertainties based on their respective data collection method. First, the comparison of validation is limited to the events and the respective extents cataloged within the databases. Second, flood maps are likely to be derived from various satellite sensors, which also impedes consistent validation across resolutions [Risling et al., 2024].

Conclusion Primary and Secondary Analysis

In sum, primary and secondary findings robustly demonstrate that diagnostic downscaling to a 15 arcsec resolution significantly enhances model performance, mainly by reducing the occurrence of false positive predictions. As GFM are increasingly relied upon to inform disaster risk management and policy decisions, the ability to refine outputs through effective downscaling methods is paramount [Ward et al., 2015]. My findings carry important implications for the use and validation of global flood models in both academic and operational contexts, as the reduction in false positives comes at the cost of an increase in false negatives. The disproportionate consequences of false positive as opposed to false negative predictions therefore necessitate careful consideration of the accuracy of GFM and substantial further research into eliminating especially false negative predictions.

Overall, the gains in accuracy are statistically robust and align well with emerging trends in high-resolution flood modeling [Wing et al., 2024, Risling et al., 2024, Bernhofen et al., 2018, Sampson et al., 2015]. The results not only validate the mechanistic advantages of diagnostic downscaling over resampling, but also underscore the necessity of high-fidelity digital elevation data to effectively capture subgrid-scale hydrodynamic and topographic variations.

Furthermore, the differential impact observed due to the resampling of the observational dataset—where a lower degradation at 15 arcsec preserves more native detail—emphasizes the challenges inherent in reconciling datasets of differing native resolutions and highlights the resulting difficulties in consistent and unambiguous validation efforts.

4.3 Performance Differs Across Biomes

A final layer of analysis was provided by evaluating model performance under differing ecological conditions, as represented by the biomes tropical rainforests, tropical savannas, flooded grasslands and deserts. The uneven distribution of flood events across biomes is problematic, as more events occur in tropical rainforests (n = 8), than in tropical savannas and flooded grasslands (each n = 4) and deserts (n = 2). This influences the reliability and generalizability of biome-specific comparisons. Furthermore, the choices of flood events are biased towards Asia, and more specifically South-to-South Asia (figure 5). However, this might be viewed as an artifact of the fact that most flood events in the Global Flood Database occur in that region. Future research should broaden and diversify the dataset to encompass a wider and more balanced array of events.

Two preliminary results emerged from this investigation.

First, variations in flood modeling performance were observed across different biomes, as the performance at both 1 arcmin and 15 arcsec was consistently highest in the tropical rainforest (median CSI about 0.1 above average). Flooded grasslands events were generally close to average, whereas tropical savannas and deserts showed the lowest scores (median CSI ca. 0.12 below average, figure 13). Widely recognized in the literature, arid conditions (desert) present difficulties in modeling and identifying surface water using remote sensing data, which might explain the poor CSI scores in these regions [Moghim et al., 2023, Bernhofen et al., 2018, Risling et al., 2024, Mester et al., 2021]. However, similar difficulties are documented for areas of dense vegetation, such as tropical rainforests, making the effective performance of CaMa-Flood in these locations during this study unexpected. Overall, my findings align with the well-established notion that the effectiveness of flood modeling varies depending on the ecological context [Moghim et al., 2023, Bernhofen et al., 2018, Risling et al., 2024, Mester et al., 2021].

Secondly, the benefits of diagnostic downscaling are not equally distributed, as demonstrated by the varying changes in performance scores across different biomes. Tropical rainforests experienced the most significant improvement in CSI—considered the most comprehensive performance metric [Bernhofen et al., 2018]. However, the origin of this improvement is counterintuitive, as the other biomes, notably tropical savannas and flooded grasslands experience more significant decreases in Bias—in other words, overestimation is reduced more markedly in these biomes at 15 arcsec. Moreover, in all biomes Hit Rate tends to decrease at 15 arcsec relative to 1 arcmin. Its decline is especially pronounced in the tropical rainforests. Consequently, the source of the superior CSI improvement observed in tropical rainforests is not immediately apparent. A closer examination of the performance score calculation logic reveals the underlying explanation, as development of CSI is inherently non-linear. For a given absolute improvement in the reduction of overestimation, the resulting increase in CSI is greater, when the baseline score is higher compared to a lower baseline. For instance, since tropical rainforests already have higher CSI scores, higher Hit Rate and lower overestimation (Bias) at 1 arcmin (overall better performance), even a smaller decrease in false alarms (overestimation) when moving to 15 arcsec leads to a larger improvement in CSI. Conversely, a lesser decrease in Hit Rate and larger improvement in Bias in tropical savannas and deserts occurs in a context where the overall error remains higher.

Limitations

This was a limited, preliminary, and, given the nature of the biome concept, necessarily superficial examination into the question of whether performance outcomes and the effects of diagnostic downscaling vary across different ecological contexts.

The results show that differences in the performance of flood models can be detected on the scale of biomes, and even that the effect of diagnostic downscaling differs between biomes. These biome-specific differences hint that underlying ecological factors might play a role in determining the efficacy of flood modeling in general and diagnostic downscaling in particular. However, while deviations were detected, the evidence does not support the conclusion that these differences are causally related to variations in ecological contexts. Even less do the results indicate any potential mechanistic pathways through which ecological factors could influence flood modeling performance and diagnostic downscaling.

Limiting the interpretability of the findings are five factors: **First**, the sample size was low in all cases. **Second**, the choice of regions was based on insufficiently rigorous criteria for selection and subsequent subsetting of events. **Third**, ecological variables were not carefully separated, forbidding any mechanistic explanation. **Fourth**, as outlined in Mester et al. [2021], the CSI score is influenced by the magnitude of the flood, which may result in higher CSI scores for larger flood events compared to those of lesser magnitude. **Fifth**, as previously observed, the improvement in CSI was mainly due to an initial high starting position.

Stated simply, although the findings suggest that both modeling and diagnostic downscaling could be improved with better ecological data, these findings are purely correlational and intertwined with other variables. It is crucial to establish more stringent criteria for event selection, such as achieving the greatest possible uniformity in event characteristics except for biome differences, ensuring events of comparable topography and extent, as well as factoring in the season. Although challenging, this approach is feasible given the availability of historical flood data in various databases, and it is essential for improving the informational value of these correlational results.

The question remains as to whether, at the biome level, this is a worthwhile avenue of investigation. Biomes are broad ecological units on continental scales, characterized by a vegetation type with a typical physiognomy. In a mechanistic sense, traits attributed to a given biome—such as vegetation cover, root density, soil permeability, and evapotranspiration rates—can be expected to influence runoff generation, infiltration, and ultimately flood behavior [Keith et al., 2020, Mucina, 2019].

However, the operationalization of traits within biome classifications is too coarse to capture the small-scale and complex interactions between landscape features and hydrological processes. For example, two regions classified under the same biome may exhibit considerable local variation in key functional traits (vegetation cover, root density, soil permeability) due to microclimatic differences, human impacts, or legacy effects from previous land uses [Keith et al., 2020, Mucina, 2019]. This variability challenges the assumption that a biome, as a discrete unit, can provide more than a coarse indicator to guide model development in ecologically diverse landscapes. However, if it were expanded to a more rigorous choice of regions and events, as described above, which are in turn more diverse and equally distributed across biomes, an investigation similar to this study could potentially be a useful guide toward subsequent closer mechanistic investigations.

A goal of global flood modeling is to achieve higher resolutions with higher fidelity, allowing low-cost and globally accessible risk management tools, useful in practical applications [Ward et al., 2015]. Should a more rigorous study support the results of this study (differences in flood modeling and diagnostic downscaling correlate to biome differences), subsequently a more nuanced approach than biome-level analysis is necessary to mechanistically improve model performance and provide more reliable flood predictions. This should involve the integration of finer-scale ecological indicators or locally calibrated parameters. For example, the ecological information used could be provided at the same spatial resolution as the flood modeling data itself. Thus, more realistic deductions on the actual effect of ecological factors on flood modeling could be achieved. For example, diagnostic downscaling is based on digital elevation models, which do not incorporate vegetation information. As vegetation has an effect on flood behavior, incorporating vegetation information into downscaling could be valuable.

In future endeavors, I recommend focusing on regions considerably smaller than those examined in this study, particularly those centered on areas that encompass elements where diagnostic downscaling could be relevant and variably influenced by features such as topography and vegetation. This could enable a detailed examination of the effects of diagnostic downscaling in these targeted regions.

Conclusion Tertiary Analysis

In conclusion, the findings suggest that flood modeling performance and the benefits of diagnostic downscaling are potentially modulated by the underlying ecological context, as evidenced by the variably distributed improvements observed across biomes. However, these findings are purely correlational and intertwined with other variables, forbidding a conclusion that these differences are causally related to variations in ecological contexts. Further biome-level research would need to establish more stringent criteria for event selection by achieving the greatest possible uniformity in event characteristics except for biome differences. More importantly, to mechanistically improve model performance and provide more reliable flood predictions, a far more nuanced approach is necessary than biome-level analysis.

4.4 Conclusion

In conclusion, this study robustly demonstrated that diagnostic downscaling markedly improves the performance of the CaMa-Flood global flood model by reducing false positive predictions, albeit at the cost of a modest increase in false negatives. The comparative analysis revealed that while both diagnostic downscaling and resampling improved model outputs, the mechanistic advantages of the former—stemming from its ability to explicitly incorporate high-resolution topographic information—yield statistically and practically superior gains. These gains are further contextualized by the observation that the resampling of observational datasets can result in a reduction of data fidelity. This emphasizes the challenges inherent in reconciling datasets of differing native resolutions, and highlights the resulting difficulties in consistent and unambiguous validation efforts.

The study provided inconclusive evidence on the distribution of performance and benefits of diagnostic downscaling across biomes. Differences were detected, though these correlational findings can not be traced to differences in ecological contexts with any confidence. The biome-analysis thus mainly highlighted the need for a more rigorous biome-analysis, and a more nuanced and spatially explicit approach to incorporating ecological contexts into future flood modeling studies.

The study's findings contain important implications for global flood modeling, both in research and practical contexts. As GFM are increasingly relied upon to inform disaster risk management and policy decisions, the ability to refine outputs through effective downscaling methods is paramount [Ward et al., 2015]. The reduction in false positives comes at the cost of an increase in false negatives. The disproportionate consequences of false positive as opposed to false negative predictions therefore necessitates careful consideration of the accuracy of GFM and substantial further research into eliminating especially false negative predictions.

Moreover, the comparison with observational flood maps from the GFD underscores the importance of aligning high-resolution model outputs with high-resolution satellite data. By resampling not only the observational data to match the model resolutions but also the model data, this study highlighted the. Thus, this study suggests that the best protocol to use for flood validation is a protocol that uses high-resolution simulated and observed data, introducing the caveat that any validation exercise should attempt to minimize resampling of datasets without additional information as much as possible. This will result in performance scores and flood maps that contain uncertainties at the level of these resolutions.

Finally, the intricacies of performance illuminated by the comparison of multiple performance scores (false-positive / false-negative trade-off; CSI / Bias improvements across biomes), each contributing separately to the overall picture, highlights the benefits of and need to incorporate multiple performance scores into validation exercises.

4.5 Outlook

While acknowledging the methodological challenges and limitations encountered, my study lays a robust foundation for future research that can build upon these findings. Looking ahead, future research should aim to broaden the scope of this investigation by incorporating multiple hydrological models and climate forcings. It should investigate the highest resolutions achievable in global flood modeling, relying on similarly higher native resolution observational datasets. Integrating flood protection measures and finer-scale ecological indicators into the modeling framework will help address existing limitations and improve the robustness of performance comparisons. Moreover, targeted studies that focus on smaller, ecologically heterogeneous regions could provide deeper insights into the interplay between topography, vegetation, diagnostic downscaling and flood dynamics, ultimately paving the way for more precise and locally relevant flood risk assessments.

Bibliography

Lorenzo Alfieri, Berny Bisselink, Francesco Dottori, Gustavo Naumann, Ad de Roo, Peter Salamon, Klaus Wyser, and Luc Feyen. Global projections of river flood risk in a warmer world. *Earth's Future*, 5(2):171–182, 2017.

Nigel W Arnell and Simon N Gosling. The impacts of climate change on river flood risk at the global scale. *Climatic Change*, 134:387–401, 2016.

Nigel W Arnell and Ben Lloyd-Hughes. The global-scale impacts of climate change on water resources and flooding under new climate and socio-economic scenarios. *Climatic change*, 122:127–140, 2014.

Peter M. Atkinson. Downscaling in remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, 22:106–114, 2013. ISSN 1569-8432. doi: 10.1016/j.jag.2012.04.012. URL `https://www.sciencedirect.com/science/article/pii/S0303243412000918`. Spatial Statistics for Mapping the Environment.

Mark V Bernhofen, Charlie Whyman, Mark A Trigg, P Andrew Sleigh, Andrew M Smith, Christopher C Sampson, Dai Yamazaki, Philip J Ward, Roberto Rudari, Florian Pappenberger, et al. A first collective validation of global fluvial flood models for major floods in nigeria and mozambique. *Environmental Research Letters*, 13(10):104007, 2018.

Seth Bryant, Guy Schumann, Heiko Apel, Heidi Kreibich, and Bruno Merz. Resolution enhancement of flood inundation grids. *Hydrology and Earth System Sciences*, 28:575–588, 2024. doi: 10.5194/hess-28-575-2024. URL `https://doi.org/10.5194/hess-28-575-2024`.

Domingos Cardoso, Peter W. Moonlight, Gustavo Ramos, Graeme Oatley, Christopher Dudley, Edeline Gagnon, Luciano Paganucci de Queiroz, R. Toby Pennington, and Tiina E. Särkinen. Defining biologically meaningful biomes through floristic, functional, and phylogenetic data. *Frontiers in Ecology and Evolution*, 9, 2021. doi: 10.3389/fevo.2021.723558. URL `https://www.frontiersin.org/articles/10.3389/fevo.2021.723558/full`.

Marco Cucchi, Graham P Weedon, Alessandro Amici, Nicolas Bellouin, Stefan Lange, Hannes Müller Schmied, Hans Hersbach, and Carlo Buontempo. Wfde5: bias-adjusted era5 reanalysis data for impact studies. *Earth System Science Data*, 12(3):2097–2120, 2020.

Eric Dinerstein, David M Olson, Douglas J Graham, Avis L Webster, Steven A Primm, Marnie P Bookbinder, and George Ledec. *A conservation assessment of the terrestrial ecoregions of Latin America and the Caribbean.* 1995.

Eric Dinerstein, David Olson, Anup Joshi, Carly Vynne, Neil D Burgess, Eric Wikramanayake, Nathan Hahn, Suzanne Palminteri, Prashant Hedao, Reed Noss, et al. An ecoregion-based approach to protecting half the terrestrial realm. *BioScience*, 67(6):534–545, 2017.

Paul A Dirmeyer, Xiang Gao, Mei Zhao, Zhichang Guo, Taikan Oki, and Naota Hanasaki. Gswp-2: Multimodel analysis and implications for our perception of the land surface. *Bulletin of the American Meteorological Society*, 87(10): 1381–1398, 2006.

Hong Xuan Do, Fang Zhao, Seth Westra, Michael Leonard, Lukas Gudmundsson, et al. Historical and future changes in global flood magnitude—evidence from a model–observation investigation. *Hydrology and Earth System Sciences*, 24: 1543–1564, 2020. doi: 10.5194/hess-24-1543-2020.

Francesco Dottori, Peter Salamon, Alessandra Bianchi, Lorenzo Alfieri, Feyera Aga Hirpa, and Luc Feyen. Development and evaluation of a framework for global flood hazard mapping. *Advances in water resources*, 94:87–102, 2016.

Francesco Dottori, Wojciech Szewczyk, Juan-Carlos Ciscar, Fang Zhao, Lorenzo Alfieri, Yukiko Hirabayashi, Alessandra Bianchi, Ignazio Mongelli, Katja Frieler, Richard A Betts, et al. Increased human and economic losses from river flooding with anthropogenic warming. *Nature Climate Change*, 8(9):781–786, 2018.

Mohammad Fereshtehpour, Mostafa Esmaeilzadeh, Reza Saleh Alipour, and Steven J Burian. Impacts of dem type and resolution on deep learning-based flood inundation mapping. *Earth Science Informatics*, 17(2):1125–1145, 2024.

Katja Frieler, Jan Volkholz, Stefan Lange, Jacob Schewe, Matthias Mengel, María del Rocío Rivas López, Christian Otto, Christopher PO Reyer, Dirk Nikolaus Karger, Johanna T Malle, et al. Scenario setup and forcing data for impact model evaluation and impact attribution within the third round of the inter-sectoral impact model intercomparison project (isimip3a). *Geoscientific Model Development*, 17(1):1–51, 2024.

Stefanie Heinicke, Jan Volkholz, Jacob Schewe, Simon N Gosling, Hannes Müller Schmied, Sandra Zimmermann, Matthias Mengel, Inga J Sauer, Peter Burek, Jinfeng Chang, et al. Global hydrological models continue to overestimate river discharge. *Environmental Research Letters*, 19(7):074005, 2024.

Michael J Hill, Miguel O Román, and Crystal B Schaaf. Biogeography and dynamics of global tropical and subtropical savannas: a spatiotemporal view. *Ecosystem function in savannas: Measurement and modeling at landscape to global scales*, 623, 2011.

Yukiko Hirabayashi, Roobavannan Mahendran, Sujan Koirala, Lisako Konoshima, Dai Yamazaki, Satoshi Watanabe, Hyungjun Kim, and Shinjiro Kanae. Global flood risk under climate change. *Nature climate change*, 3(9):816–821, 2013.

Chris B Joyce, Matthew Simpson, and Michelle Casanova. Future wet grasslands: ecological implications of climate change. *Ecosystem Health and Sustainability*, 2(9):e01240, 2016.

David A Keith, Jose R Ferrer-Paris, Emily Nicholson, and Richard T Kingsford. Iucn global ecosystem typology 2.0. *Descriptive profiles for biomes and ecosystem functional groups. IUCN, Gland*, 2020.

Hyungjun Kim. Global soil wetness project phase 3 atmospheric boundary conditions (experiment 1). *(No Title)*, 2017.

Elco E Koks, Brenden Jongman, Trond G Husby, and Wouter JW Botzen. Combining hazard, exposure and social vulnerability to provide lessons for flood risk management. *Environmental science & policy*, 47:42–52, 2015.

Stefan Lange. Wfde5 over land merged with era5 over the ocean (w5e5). *(No Title)*, 2019.

Roberto Lázaro, Cayetana Gascón, and Consuelo Rubio. Runoff and soil loss in biocrusts and physical crusts from the tabernas desert (southeast spain) according to rainfall intensity. *Frontiers in Microbiology*, 14:1171096, 2023.

David Adams Leeming. *World mythology: A very short introduction*, volume 716. Oxford University Press, 2022.

Tian Liu, Peijun Shi, and Jian Fang. Spatiotemporal variation in global floods with different affected areas and the contribution of influencing factors to flood-induced mortality (1985–2019). *Natural Hazards*, 111(3):2601–2625, 2022.

Benedikt Mester, Sven Norman Willner, Katja Frieler, and Jacob Schewe. Evaluation of river flood extent simulated with multiple global hydrological models and climate forcings. *Environmental Research Letters*, 16(9):094010, 2021.

Sanaz Moghim, Mohammad Ahmadi Gharehtoragh, and Ammar Safaie. Performance of the flood models in different topographies. *Journal of Hydrology*, 620: 129446, 2023.

Ladislav Mucina. Biome: evolution of a crucial ecological and biogeographical concept. *New Phytologist*, 222(1):97–114, 2019.

Hannes Müller Schmied, Tim Trautmann, Sebastian Ackermann, Denise Cáceres, Martina Flörke, Helena Gerdener, Ellen Kynast, Thedini Asali Peiris, Leonie Schiebener, Maike Schumacher, et al. The global water resources and use model watergap v2. 2e: description and evaluation of modifications and new features. *Geoscientific Model Development Discussions*, 2023:1–46, 2023.

N. Najibi and N. Devineni. Recent trends in the frequency and duration of global floods. *Earth System Dynamics*, 9(2):757–783, 2018. doi: 10.5194/esd-9-757-2018. URL `https://esd.copernicus.org/articles/9/757/2018/`.

David M Olson, Eric Dinerstein, Eric D Wikramanayake, Neil D Burgess, George VN Powell, Emma C Underwood, Jennifer A D'amico, Illanga Itoua, Holly E Strand, John C Morrison, et al. Terrestrial ecoregions of the world: A new map of life on earth: A new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. *BioScience*, 51(11):933–938, 2001.

Jean-François Pekel, Andrew Cottam, Noel Gorelick, and Alan S Belward. High-resolution mapping of global surface water and its long-term changes. *Nature*, 540(7633):418–422, 2016.

Adnan Rajib, Qianjin Zheng, Charles R Lane, Heather E Golden, Jay R Christensen, Itohaosa I Isibor, and Kris Johnson. Human alterations of the global floodplains 1992–2019. *Scientific Data*, 10(1):499, 2023.

Andy Reisinger, Mark Howden, Carolina Vera, et al. The concept of risk in the ipcc sixth assessment report: A summary of cross-working group discussions, 2020. URL `https://apps.ipcc.ch/glossary/`.

Axel Risling, Sara Lindersson, and Luigia Brandimarte. A comparison of global flood models using sentinel-1 and a change detection approach. *Natural Hazards*, pages 1–20, 2024.

Christopher C Sampson, Andrew M Smith, Paul D Bates, Jeffrey C Neal, Lorenzo Alfieri, and Jim E Freer. A high-resolution global flood hazard model. *Water resources research*, 51(9):7358–7381, 2015.

Inga J Sauer, Benedikt Mester, Katja Frieler, Sandra Zimmermann, Jacob Schewe, and Christian Otto. Limited progress in global reduction of vulnerability to flood impacts over the past two decades. *Communications Earth & Environment*, 5 (1):239, 2024.

Louise Slater, Gabriele Villarini, Stacey Archfield, Daniel Faulkner, Rob Lamb, Ali Khouakhi, and Jian Yin. Global changes in 20-year, 50-year, and 100-year river floods. *Geophysical Research Letters*, 48:e2020GL091824, 2021.

Beth Tellman, Jonathan A Sullivan, Catherine Kuhn, Albert J Kettner, Colin S Doyle, G Robert Brakenridge, Tyler A Erickson, and Daniel A Slayback. Satellite imaging reveals increased proportion of population exposed to floods. *Nature*, 596(7870):80–86, 2021.

UNDRR and CRED. The human cost of disasters: An overview of the last 20 years (2000-2019). October 2020.

Ágnes Vári, Zsolt Kozma, Beáta Pataki, Zsolt Jolánkai, Máté Kardos, Bence Decsi, Zsolt Pinke, Géza Jolánkai, László Pásztor, Sophie Condé, et al. Disentangling the ecosystem service 'flood regulation': Mechanisms and relevant ecosystem condition characteristics. *Ambio*, 51(8):1855–1870, 2022.

Philip J Ward, Brenden Jongman, Peter Salamon, Alanna Simpson, Paul Bates, Tom De Groeve, Sanne Muis, Erin Coughlan De Perez, Roberto Rudari, Mark A Trigg, et al. Usefulness and limitations of global flood risk models. *Nature Climate Change*, 5(8):712–715, 2015.

Philip J Ward, Veit Blauhut, Nadia Bloemendaal, James E Daniell, Marleen C de Ruiter, Melanie J Duncan, Robert Emberson, Susanna F Jenkins, Dalia Kirschbaum, Michael Kunz, et al. Natural hazard risk assessments at the global scale. *Natural Hazards and Earth System Sciences*, 20(4):1069–1096, 2020.

Daniel S. Wilks. *Statistical Methods in the Atmospheric Sciences*, volume 91 of *International Geophysics Series*. Academic Press, Amsterdam; Boston, 2nd edition, 2006. ISBN 9780127519661.

Oliver EJ Wing, Paul D Bates, Niall D Quinn, James TS Savage, Peter F Uhe, Anthony Cooper, Thomas P Collings, Nans Addor, Natalie S Lord, Simbi Hatchard, et al. A 30 m global flood inundation model for any climate scenario. *Water Resources Research*, 60(8):e2023WR036460, 2024.

WMO. *Atlas of Mortality and Economic Losses from Weather, Climate and Water Extremes (1970–2019).* World Meteorological Organization (WMO), Geneva, Switzerland, 2021. ISBN 978-92-63-11267-5.

Dai Yamazaki, Shinjiro Kanae, Hyungjun Kim, and Taikan Oki. A physically based description of floodplain inundation dynamics in a global river routing model. *Water Resources Research*, 47(4), 2011.

Dai Yamazaki, Daiki Ikeshima, Jeison Sosa, Paul D Bates, George H Allen, and Tamlin M Pavelsky. Merit hydro: A high-resolution global hydrography map based on latest topography dataset. *Water Resources Research*, 55(6):5053–5073, 2019.