Judgment and decision strategies used by weather scientists in southeast Asia to classify impact severity

Xiaoxiao Niu[ab], Henrik Singmann[b], Faye Wyatt[c], Agie W Putra[d], Azlai Taat[e], Jehan S Panti[f], Lam Hoang[g], Lorenzo A Moron[f], Sazali Osman[h], Riefda Novikarany[d], Diep Quang Tran[g], Rebecca Beckett[c], Adam JL Harris[b]

[a] **College of Management**, **Shenzhen University**, 1066 Xueyuan Boulevard, Nanshan, Shenzhen, Guangdong Province, 518060, China

[b] **Department of Experimental Psychology**, **University College London**, 26 Bedford Way, London, WCH1 0AP, United Kingdom

[c] **Met Office**, FitzRoy Road, Exeter, Devon, EX1 3PB, United Kingdom

[d] **Public Weather Services Center, Indonesian Agency for Meteorology, Climatology and Geophysics**, Jakarta 10720, Indonesia

[e] **MetMalaysia,** Jabatan Meteorologi Malaysia Jalan Sultan, 46200 Petaling Jaya, Selangor, Malaysia

[f] **Department of Science and Technology, Philippine Atmospheric, Geophysical and Astronomical Services Administration**, PAGASA Science Garden Complex, BIR Road, Brgy. Central, Quezon City, Metro Manila 1100, Philippines

[g] **Viet Nam National Center of Hydro-Meteorology Forecasting,** 8 Phao Dai Lang street, Lang Thuong, Dong Da, Ha Noi, Vietnam

[h] **National Flood Forecasting and Warning Centre, Department of Irrigation and Drainage Malaysia,** Km 7, Jalan Ampang, 68000 Kuala Lumpur, Malaysia

Author Contributions

AJLH conceived the study; AJLH, XN and HS designed the study with input from all authors; XN and AJLH ran the qualitative focus group discussions; FW compiled the online database; HS used the database to develop the study materials; XN programmed the study; RN and DQT were responsible for the translation of survey into Bahasa and Vietnamese. AP was responsible for data collection in Indonesia; AT and SO were responsible for data collection in Malaysia; LH was responsible for data collection in the Vietnam; LAM and JSP were responsible for data collection in the Philippines; XN analysed the data; AJLH and HS advised on analysis decisions; XN wrote the manuscript; AJLH and HS edited the manuscript; RB co-ordinated the international collaboration. All authors reviewed and commented on the manuscript.

Abstract

Impact-based weather forecasting requires forecasters to predict what weather might *do* (impact information), rather than solely what weather might *be* (meteorological information). In a collaboration between the UK Met Office, UK psychologists, and weather scientists in Indonesia, Malaysia, The Philippines, and Vietnam, the present study employed Judgment Analysis and decision strategy comparisons to better understand weather scientists' impact severity judgments. In the Judgment Analysis Task, weather scientists (from Indonesia, Malaysia, the Philippines, and Vietnam) made numerical and categorical severity judgments for 70 hypothetical heavy rainfall events, each described via six impacts (e.g., number of deaths, number of people affected). The hypothetical impacts were generated from a multivariate distribution estimated from a distribution of real rainfall events. Subsequently, participants provided categorical severity classifications for a list of impact values for each type of impact (Threshold Identification Task) to aid the identification of decision strategies. In all four countries, weather scientists' severity judgments were best predicted by incorporating all six impacts via a compensatory judgment strategy. However, considerable individual differences were identified in the weights assigned to the different impacts and in the identified thresholds for each impact's categorical severity classification. To improve impact-based forecasting, meteorological agencies should seek to enhance consistency among forecasters.

Keywords: Compensatory judgment strategy; Judgment Analysis; Impact-based forecasts; Impact-based weather warnings; severity judgments; threshold identification

1. Introduction

Mitigating costs and damages associated with extreme weather events is a global priority (United Nations, 2015). Results from the Global Risks Perception Survey (World Economic Forum, 2023) ranked such events as the second most severe risk facing the world (after the Cost-of-living crisis) in the next two years (and third in the next 10 years). Extreme weather refers to weather events that are markedly different from average or usual weather patterns, and includes events such as storms, flash-floods, strong winds, heatwaves and droughts (Haryanto et al., 2020). Extreme weather conveys significant negative impacts on populations, natural resources, and economic growth.

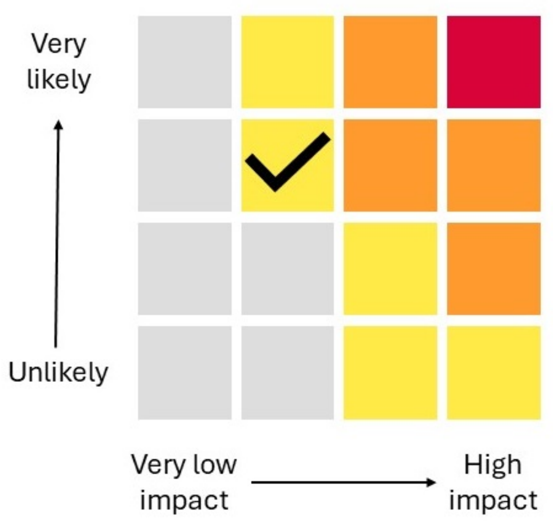
Southeast Asia is a region particularly affected by extreme weather. In Indonesia, between 1998 and 2018, extreme weather events included 8,814 floods, 5,969 heavy wind/storms, 4,946 landslides, 1,872 droughts, and 13 tsunamis (Haryanto et al., 2020). In the last two decades, Malaysia has experienced 51 extreme weather events and earthquakes, with 281 deaths, over 3 million people affected, and US $2 billion losses (Centre for Excellence in Disaster Management and Humanitarian Assistance, 2019). Over the past 20 years, extreme weather events in Vietnam have caused more than 13,000 deaths, and property damage in excess of US $6.4 billion (Global Facility for Disaster Reduction and Recovery, 2018). The Philippines is one of the most disaster-prone countries in the world and experiences an average of 20 typhoons every year (Global Facility for Disaster Reduction and Recovery, 2016). Profound psychological impacts caused by extreme weather events such as trauma, anxiety, depression are also documented in Southeast Asia (Patwary et al., 2024). The design and utilization of an effective warning system is one approach to mitigate the damages caused by such hazards.

1.1 The development of Impact Based Warnings

One approach to improve the effectiveness of weather warnings in mitigating impacts is the development of Impact-Based Warnings (IBWs) (WMO, 2015). Incorporating *impact* information into forecasts and warnings is thought to facilitate appropriate protective actions (e.g., Uccellini & Hoeve, 2019). IBWs can be well-illustrated through the Warning Impact Matrix (Figure 1, Met Office, 2023), which is employed by many countries for the development and communication of IBWs. A warning on this risk matrix conveys the severity of *impacts* the weather may cause (Very Low/ Minimal, Low/Minor, Medium/ Significant, and High/ Severe) and the *likelihood* of those impacts occurring (Very Low, Low, Medium, and High). The combination of impact severity and likelihood defines a warning level colour. Yellow and Amber warnings represent a range of potential impacts and their likelihood; a Red warning indicates that dangerous weather is expected with a high likelihood of severe impacts. The impact-based forecasting (IBF) paradigm is accepted and used by many national meteorological and hydrological services around the world (WMO, 2021b; Yu et al., 2022), with research to develop such capacities in a number of other locations (e.g., Beckett & Hartley, 2020; Harrison et al., 2021; Jenkins et al., 2022; Kaltenberger et al., 2020; Mitheu et al., 2023; Singhal et al., 2022).

Figure 1

Met Office (2023) Warning Impact Matrix



The move to IBWs clearly provides a requirement to forecast weather *impacts*, as well as (just) the weather. Consequently, algorithmic calculations of expected impacts have received recent research attention, including successful implementations of impact models for vehicles overturning (Hemingway & Robbins, 2020), building damages (Röösli et al., 2021; Wei et al., 2018), electrical infrastructure damages (Wilkinson et al., 2022), critical facility damages (Taramelli et al., 2015), and agricultural losses (Chau et al., 2015).

1.2 Impact severity classification

In addition to predicting specific impacts, successful IBW requires matching expected impacts to user-agreed overall severity levels (as in Figure 1). This classification of impacts into overall severity levels can be achieved through comparison with impact tables that have been developed for just this purpose within countries employing IBF. Impact tables comprise a collection of descriptive information of possible impacts from a variety of different sectors (e.g., Population, Transport, Education), classified according to severity level (e.g., Minimal, Minor, Significant and Severe). Impact tables are typically developed and refined through discussion between forecasters and key stakeholders (UN ESCAP, 2021; Met Office, 2017). Extant impact tables list various types of impacts in relation to different hazards. Consequently, in theory, forecasted impacts can be 'looked up' in these Impact tables, so as to define an appropriate severity level. In an online experiment with individual forecasters, Jenkins et al. (2022) observed good alignment between (hypothetical) impact-based warnings issued by forecasters, with those implied by the original impact tables in Indonesia, Malaysia, and the Philippines.

Jenkins et al. (2022) presented participants with a single expected impact, with an identical description to that found in their country's impact table. There are, however, challenges associated with the use of impact tables for severity classification. A key one is the mismatch between the qualitative format of impact tables (e.g., "wider-scale and prolonged disruption to daily life and services") and the often quantitative outputs of impact forecast algorithms. Developers of these algorithms have taken a variety of approaches to attempt to link their quantitative outputs with an overall severity classification (Aldridge et al., 2020; Aldridge et al., 2016; Cole et al., 2016; Moore et al., 2015; Speight et al., 2018). Aldridge et al. (2016), for example, classified 40-199 lives in danger as minor, 200-299 lives in danger as significant, and 300+ lives in danger as severe. Wyatt et al. (2023) designated severity classifications via octile breaks. In some other studies, researchers used both quantitative thresholds and descriptive classifications for impact severity assessment (Robbins & Titley, 2018; Spruce et al., 2021). All the above approaches are inherently sensible. It is not, however, known how the approaches match users' (both forecasters & forecast recipients) interpretations of how numerical impact forecasts map onto severity classifications.

Moreover, extant impact tables only match the level of *individual* sector impacts to a severity classification. We are aware of no work investigating how forecasted impacts on different sectors (e.g., agriculture & travel) are, or should be, combined in an overall severity classification.

In addition to its importance for informing appropriate severity classifications for the issuance of IBWs, understanding how quantitative impacts are combined in an overall severity classification judgment is essential for the *evaluation* of IBWs. Evaluation requires a comparison between the warning level issued, and the impacts of a weather event that are actually observed. The impacts associated with extreme weather events are commonly recorded and reported in a numerical format, such as the number of people dead, and the number of livestock killed (AHA Center, 2023; ECHO, 2023; OCHA, 2022; WMO, 2021a). Notwithstanding that certain agencies prefer to report particular types of impacts or hazard - a reporting bias that can be reduced by combining observations across multiple information sources (Wyatt et al., 2023) - these numerical records offer an appealing way to evaluate IBF by comparing its forecasted impacts with observed outcomes. To do so, however, requires that the observed quantitative impact information reported by these sources can be mapped back to an overall (usually categorical – see Figure 1) severity forecast issued as part of an IBW.

Identifying how numerical information about weather impacts is combined in overall severity classifications can therefore inform both forecast operations and IBF evaluation. From an operational perspective, identifying the strategies that forecasters actually use (a descriptive approach) does not, prima facie, imply that those are the ones that *should* be used (a normative question). Such identification does, however, enable subsequent scrutiny. Senior forecasters and disaster managers can subsequently determine whether the strategies used are appropriate, and employ training methods where this is thought not to be the case (see e.g., Cooksey, 1996, for examples and best practice of such applications of Judgment analysis). Even without such scrutiny, a descriptive approach can identify the degree of (in)consistency between forecasters.

We are aware of only one study that has sought to identify the quantitative impacts associated with different severity classifications. Sai et al. (2018) asked farmers to recall floods that had ‘minor’, ‘significant’ and ‘severe’ impacts. They found that events that caused less than 10% crop damages were rated as minor events, events that caused 60% - 80% crop damages and loss of livestock were significant events, and events that caused 80%-100% crop damages and diffuse loss of livestock were severe events. This study is clearly limited in scope, and it is also possible that when farmers defined the severity levels of the scenarios, they not only focused on the impact of crop and livestock, but also considered impacts from other sectors.

1.3 The current study

The primary aim of the current study is to understand how weather scientists[1] form overall severity classifications from sets of numerical impact values. Specifically, a Judgment Analysis approach (Cooksey, 1996) is used to reveal how weather scientists utilize, weight, and combine different numerical impact information into an overall severity judgment. Commonalities or inconsistencies between weather scientists can be identified. Where there is good consistency between weather scientists, the results from this study may be used to assign severity thresholds to observed numerical impacts in disaster databases or impact reports, allowing for the subsequent evaluation of IBWs.

Judgment Analysis can identify if weather scientists rely on a small sub-sample of impact types (e.g., only number of people affected) to make their severity classifications. Judgment and decision strategies relying on single cues, or a sequential consideration of single cues (e.g., following a Take the best approach - Gigerenzer & Goldstein, 1996; Gigerenzer et al., 2011; Gigerenzer & Todd, 1999), can be classed as non-compensatory. Alternatively, some judgment and decision strategies involve consideration of multiple cues at once. For example, a mean strategy (whereby the final severity classification is a [weighted] mean of the severity level implied by each individual impact) can be considered a compensatory strategy, as a low value for one impact can be ‘compensated’ for by a high value of another. As part of our aim of determining how weather scientists combined multiple pieces of quantitative impact information to form an overall severity judgment, we sought to

[1] Scientists from Southeast Asia meteorological agencies (including, but not limited to operational forecasters) were included as potential participants, given that their knowledge is driving IBF in these countries.

ascertain the degree to which they engaged in a compensatory or non-compensatory judgment process (Dhami & Harries, 2001; Foerster, 1979).

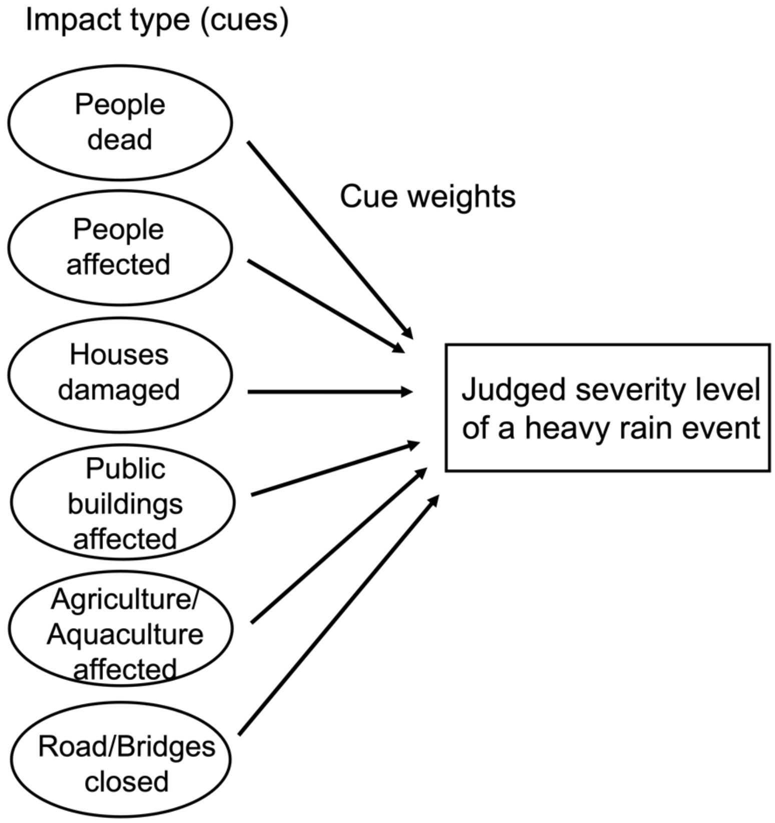
Whilst Judgment Analysis can provide clues to the compensatory vs. non-compensatory question (i.e., if only 'number of people affected' is a significant predictor of overall severity judgments this points towards a non-compensatory strategy), we wanted to extend this to consider the possibility that the impacts weather scientists focus on (in a non-compensatory process) depend on the severity level of those individual impacts. For example, an intuitively plausible decision process is that weather scientists judge overall severity level according to the highest severity level implied by any individual impact type. Indeed, Aldridge et al. (2016) adopted just this approach when determining the overall severity level for each 1km grid cell of a focal area, across all contributing impact criteria. Meyer et al. (2007) also argued that such a disjunctive approach is more appropriate for flood risk analysis than other approaches (e.g., a conjunctive approach whereby a severe severity classification is only made if all impact types are individually severe). These are just two of a number of possible algorithmic strategies of aggregating individual impacts that forecasters might use to make an overall assessment (for overviews of such strategies see Malczewski, 1999; Malczewski & Rinner, 2015).

1.3.1. Judgment Analysis

Judgment Analysis relies on multiple regression to investigate how people weight different pieces of information when making judgments. We apply this approach to examine how weather scientists' severity judgments are influenced by the quantitative levels of different impacts of a weather event (Figure 2).

Figure 2

Overall severity judgments as a function of individual impact levels


Impact type (cues)
People dead
People affected
Houses damaged
Public buildings affected
Agriculture/ Aquaculture affected
Road/Bridges closed
Cue weights
Judged severity level of a heavy rain event

In our Judgment Analysis task, weather scientists make severity judgments for a series of hypothetical heavy rainfall events. Either hypothetical events or real events can be employed in Judgment Analysis, and each approach has its advantages and disadvantages (Cooksey, 1996; Dhami et al., 2004). A key advantage of using hypothetical events is that researchers can remove correlations between cues and utilise fully factorial designs. However, such correlations exist in the real world, and this is certainly true in the case of weather impacts. There is therefore a trade-off between multicollinearity and ecological validity. Usually the key problem posed by multicollinearity is that estimates of the unique predictiveness of any particular cue (e.g., impact type) are specific to the particular correlational structure in the Judgment Analysis task - the estimates reflect the unique contribution of each predictor (i.e., after accounting for the contribution of the other predictors). In the current study, this can be viewed as a desirable property *if* the correlations in our task approximate those for real-world weather events. Our aim is to understand the structure of weather scientists' severity judgments for real-world weather forecasts, which necessarily have high degrees of multicollinearity (i.e., less extreme weather hazards tend to lead to lower consequences across all impact types and more extreme weather hazards lead to greater consequences across all impact types). Consequently, we sought to maximise the ecological validity of our task structure in the current study. This was achieved through creating a task structure in which the correlations between cue values approximated those observed for real weather events, as identified from databases reporting the impacts of real weather events.

1.4 Previous Findings

The dominant result in the analysis of expert judgment is that experts typically rely on a small sub-sample of cues to form judgments (e.g., using 5 out of 10 presented). Such a result is observed across domains, including medical diagnosis and referral (Saintonge et al., 1988; Harries et al., 1996; Baker & Thompson, 2012), accounting and finance (Ettenson et al., 1987; Kuo & Liang, 200), and Education (Browne & Gillis, 1982). Such a result is not, however, universal. With a tabular presentation of symptoms, White et al. (2018) found palliative care professionals utilised six out of seven in predicting imminent death, and eight out of fourteen cues were significantly weighted when bankers rated loan applicants (Wilsted et al., 1975).

Judgment analysis has also been applied to weather forecasting, where Stewart et al. (1989) argued that regression models provided good descriptions of meterologists' forecasts. Mirroring the majority of research in this tradition, meteorologists have, however, typically been found to rely on a sub-sample of the cues provided (e.g., 2-4 out of 12-24 cues; Stewart et al., 1997; see also Stewart et al., 1989; Stewart et al., 1992). These previous studies have focussed solely on the prediction of weather phenomena (e.g, precipitation likelihood, maximum temperature). In the current study, we specifically investigate weather scientists' judgments of *impact severity*. Due to the novelty of this research, we do not make specific predictions, but note that we expect the results to be informative for understanding how such judgments are made, and for the subsequent evaluation of IBWs.

A considerable amount of research has been published on impact-based forecasting recently (e.g., Boult et al., 2022; Mitheu et al. 2023; Nkiaka et al. 2020; Potter, et al., 2021;

Sai et al., 2018; Silvestro et al., 2019). This is the first study to investigate how weather scientists' severity judgments relate to multiple (quantitatively expressed) impacts. We aim to understand how weather scientists from four Southeast Asian countries (Indonesia, Malaysia, The Philippines, Vietnam)[2] combine information from six impact types to make an overall severity judgment. We employ a mixed-methods approach, whereby our quantitative empirical study was informed by analyses of extant weather-event databases and qualitative discussions with in-region forecasters to ensure ecological validity and relevance of the cues included.

2. Methods

We conducted the same study in Indonesia, Malaysia, the Philippines, and Vietnam. We report the general methodology once, and highlight differences between countries where appropriate. The method, but not the analysis plan, was pre-registered. Pre-registration, materials and analysis code are available at https://osf.io/84scv/?view_only=fd5c264109374a48ac8fe7558fad999a.

The study included two tasks, the Weather Event Judgment Analysis Task and the Individual Impact Threshold Identification Task (hereafter referred to as 'Judgment Analysis Task' and 'Impact Threshold Task'). The study was presented in English in Malaysia and the Philippines. It was translated into Bahasa in Indonesia, and Vietnamese in Vietnam.

2.1 Participants

Weather scientists working in the four countries were recruited through online survey links, distributed via email. Reminder emails were sent weekly throughout the data collection period (see Table 1), except during local holidays. Participants volunteered for this study and were not monetarily reimbursed. Ethical approval was granted from the Departmental Ethics Chair for Experimental Psychology (University College London). Pre-registered

[2] The organizations which are responsible for impact-based forecasting in each country are: the Indonesian Agency for Meteorology, Climatology and Geophysics (BMKG) in Indonesia, the Philippine Atmospheric, Geophysical and Astronomical Services Administration (PAGASA) in Philippines, Viet Nam National Center of Hydro-Meteorology Forecasting (NCHMF) in Vietnam, and the Department of Irrigation and Drainage Malaysia and Met Malaysia in Malaysia. Fifty-six participants from Indonesia are students from State College of Meteorology Climatology and Geophysics. They are familiar with IBF through their studies, but do not engage in daily practice.

exclusion criteria included: (1) unfinished survey responses, (2) respondents with missing values on an item, and (3) responses that did not respect monotonicity in the Impact Threshold Task[3]. Only the third criterion led to any exclusions (6 participants from Vietnam). Data from 278 participants were subsequently included in the analysis.

Table 1

Participant information for four partner countries

Country		Indonesia	Malaysia	Philippines	Vietnam
Full completion		105	45	33	95
Location	Headquarter	2	35	20	55
	District	103	10	13	40

Organization		BMKG(49)
		STMKG(56)
			DID (13)
			MetMalaysia
			Other (2)
				PAGASA	NCHMF
					Other (2)

Experience with IBF	Little or no experience	28.6%	37.8%	33.3%	36.6%
	I have received training on it	26.7%	15.6%	30.3%	21.8%
	Some experience	41.0%	46.7%	33.3%	39.6%
	A lot of experience	3.8%	0.0%	3.0%	2.0%
Experience with IBF risk matrix	Little or no experience	27.6%	48.9%	42.4%	42.6%
	I have received training on it	30.5%	15.6%	27.3%	17.8%
	Some experience	39.0%	33.3%	30.3%	38.6%
	A lot of experience	2.9%	2.2%	0.0%	1.0%
Use IBF	Don't use IBF	27.6%	68.9%	87.9%	52.5%
	Use IBF	72.4%	31.1%	12.1%	47.5%
Data collection period[4]		December 1st 2022 - January 17th 2023	December 1st 2022 - January 31st 2023	December 2nd 2022 - January 31st 2023	December 15th 2022 - January 31st 2023

[3] In the Impact Threshold Task, participants had to provide categorical severity ratings (Minimal, Minor, Significant, and Severe) for several numerical impacts, separately for each impact dimension (e.g., first for number of deaths then for number affected). Importantly, for each impact dimension the numbers that had to be rated increased in ascending order (e.g., first 1 death, then 2 deaths, ...). There should therefore be a corresponding monotonic increase in severity ratings (first Minimal, then Minor, etc.). Any deviations from strict monotonicity led to participant exclusion. For example, one participant's data were excluded since they rated a heavy rainfall event that caused 3 houses damaged as a significant severity event but rated a heavy rainfall event that caused 7,347 houses damaged as a minor severity event.

[4] We pre-registered that we would collect data within the timeframe of one calendar month and aim to collect 20 participants in each country. Given the difficulties with obtaining the minimum number of participants after four working weeks in Malaysia, Philippines and Vietnam, we subsequently extended the data collection period for them.

2.2 Materials

2.2.1 Identification of impacts for heavy rainfall events

Real weather events and corresponding impacts were collected from seven online databases compiled into one (hereafter referred to as 'the database') (Wyatt et al., 2023). The majority of weather events recorded in the database are precipitation related, with the most common events being floods and flash floods. Additionally, in-country authors have consistently identified heavy rainfall events as the most relevant for their forecasters. These concerns led us to focus the current study on heavy rainfall events. In addition to information from the database, a qualitative study was conducted to help finalise the impacts for heavy rainfall events. The six impact types used in the Judgment Analysis Task (Figure 2) were selected based on the data available in the database, in conjunction with the results of four qualitative surveys and four discussion groups conducted in August 2022 (one qualitative survey and one focus discussion for each country).

2.2.1.1. Database analysis

The database (for details see Wyatt et al., 2023) included five global impact data sources and two regional impact data sources covering Southeast Asia. Nineteen impact types were identified in the database (e.g., the number of reported deaths, the number of hospitals or health centres affected, the total economic losses reported).

2.2.1.2. Qualitative survey and Focussed discussions

A qualitative survey was distributed to a small number of key weather scientists to obtain a complete list of impact types they consider for heavy rainfall events. Nineteen survey responses were collected from four countries, including 4 responses from BMKG Indonesia answered by author R.N. and 3 colleagues, 4 responses from PAGASA Philippines answered by author L.A.M., J.S.P. and 2 colleagues, 5 responses from NCHMF Vietnam answered by author L. H. and 4 colleagues, and 6 responses from Malaysia answered by authors A.T., S.O., and 4 colleagues from Met Malaysia. In the survey, participants were asked: "When you think of Heavy Rain, what impacts would you consider in determining the severity of a heavy rain event". Fifty-five unique impacts of heavy rainfall events were identified, and a list of those impacts was sent back to the participants before focus discussion (see Appendix 1).

It was necessary to reduce the number of impacts to be included in the quantitative survey, as each additional impact requires a minimum of five additional events (and more where there are high correlations between cues) to enable reliable estimates of each impact's regression weight (Cooksey, 1996). It was highly desirable to keep the survey as short as possible in order to encourage participation from as many of our intended participants as possible. To identify the most critical impacts when weather scientists judge the severity level of the heavy rainfall events, separate online focus discussions (each lasting about 2 hours) were conducted with each country's representatives. Although aiming to recruit the same experts who had participated in the initial survey to focus discussions, due to the unavailability of some experts, two experts from BMKG Indonesia were replaced by two new experts; one expert from NCHMF Vietnam was lost; one additional expert from DID Malaysia was recruited. Approximately one week prior to the focus discussions, a summary of the impacts listed in the qualitative survey was provided back to the experts. During each discussion, the experts discussed: how multiple impacts can be grouped together, whether any key impacts were missing if only focusing on six impacts that were commonly mentioned in both qualitative survey and the database, how to distinguish different impacts, how spatial and temporal factors relate to the six impacts, which format is easiest to answer questions (text vs table), who are the people that receive impact forecast information and make severity judgments[5]. Consequently, six key impacts were identified: People dead, People affected (e.g., injured, displaced, evacuated), Houses damaged or destroyed, Public buildings affected (e.g., schools, hospitals, government or religious buildings), Agriculture/aquaculture affected (hectares), and Road sections and/or bridges closed. Within these focus discussions, our experts confirmed that there were no critical impacts missing if we focussed on these six.

[5] Slides used to guide the focussed discussions are available at https://osf.io/84scv/?view_only=fd5c264109374a48ac8fe7558fad999a.

2.2.2 Identifying impact values

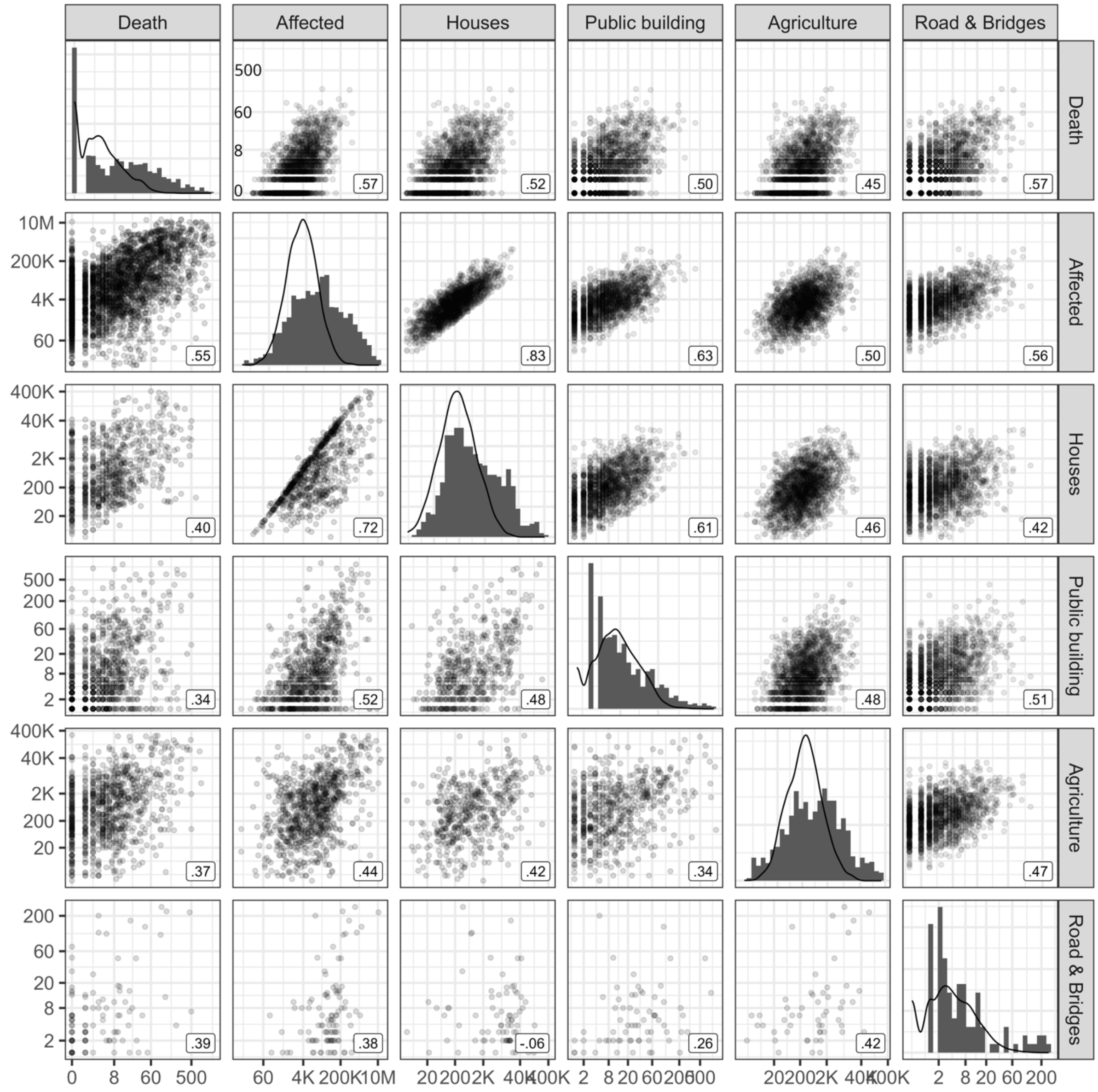
The levels of the six impacts for all presented hypothetical heavy rainfall events were simulated from a multivariate normal distribution that was estimated from real heavy rainfall events present in the database. This ensured that all hypothetical events used in this study had high ecological validity, as they were statistically similar to real heavy rainfall events. The statistical similarity included the collinearity among different impacts that is typical for real rainfall events. This similarity is visualised in Figure 3, which shows the pair-wise joint distributions (co-occurrences) of the real weather impacts in the lower triangle, and the pair-wise joint distributions of the hypothetical weather impacts in the upper triangle. The univariate distribution of each variable is shown in the diagonal, with the grey histogram showing the real impacts and the density estimates (solid lines) showing the hypothetical impacts.

Full technical details of how the task events were obtained based on the database can be found in Appendix 2. Worthy of note here, though, is that the distributions of all impacts were extremely positively skewed, with most having some extreme outliers. This reflects the fact that the most catastrophic weather events are (close to) one-of-a-kind events, with many more low impact events occurring across the world. To address this property of the distributions, and the fact that there are many events with zero deaths in the database, we winsorised each impact type distribution at the 99th percentile, and performed a log1P (defined as ln[1+x]) transformation. The multivariate normal distribution was estimated to the impact events on the transformed scale, which is also the scale shown in Figure 3. Whereas the Judgment Analysis (as well as the other statistical analyses) were also performed on the transformed scale, the events shown to participants in the study (see Figure 4) were back-transformed from the log1P scale ($e^x - 1$). We also checked the hypothetical events for patterns that were implausible (specifically, in 67 of the hypothetical 2,000 events the number of affected people was less than two times the number of affected houses; likewise, in 21 events there were more affected public buildings than affected houses). We removed these implausible events from the list of possible events. Each weather event presented to participants in the Judgment Analysis Task was randomly sampled (with replacement across participants) from the remaining 1,912 hypothetical events created from the database. One additional constraint to the sampling was that 70%

(49) of events presented to each participant should have no deaths associated with it. This decision followed an intuition (not subsequently borne out in the results, but reflected in some focus group discussions) that a consideration of deaths would overshadow any consideration of other impact types. Consequently, we wanted to ensure that we had sufficient events without any deaths observed to be able to measure the influence of all impact types.

Figure 3

Multivariate Distribution of Real and Hypothetical Weather Events



Note. The lower triangle (panels below the diagonal) shows the pair-wise distributions of real weather events and the upper triangle (panels above the diagonal) shows the pair-wise distributions of hypothetical (i.e., simulated) weather events. In these panels the values shown are the numbers of the corresponding impacts (e.g., number of deaths on the x-axis in the first column and y-axis of the first row). The value in each lower right corner is the correlation between each pair of variables shown in each panel. The diagonal panels show the univariate distribution of each impact where the histogram (in grey) shows the distribution of real events and the density estimate (the black line) the distribution of the hypothetical events. Because the y-axes of the diagonal panels showing the density do not match the y-axes of the pairwise plots showing the pairwise data, the y-axis labels for the first row are shown only in the second panel. Because of the strong right skew of each univariate distribution, all variables are log1P transformed ($x_{\text{transformed}} = \ln(1 + x)$), but axes values are given on the original (i.e., event) scale. The lower triangle of Figure 3 shows the multivariate distribution of events in a pair-wise manner. This shows that there is a substantive positive correlation among all (log1P transformed) impacts. The only exception is the correlation between houses and roads & bridges, which we assume is a consequence of the limited data for these two variables. A good match between the correlations in the real and hypothetical weather events is shown if the pair-wise distributions in the lower triangle resemble a transposed version of the corresponding distributions in the upper triangle. This is generally observed in this figure across pairs with the pair-wise correlations for the hypothetical events being somewhat larger than the correlations for the real events.

2.3 Task Design

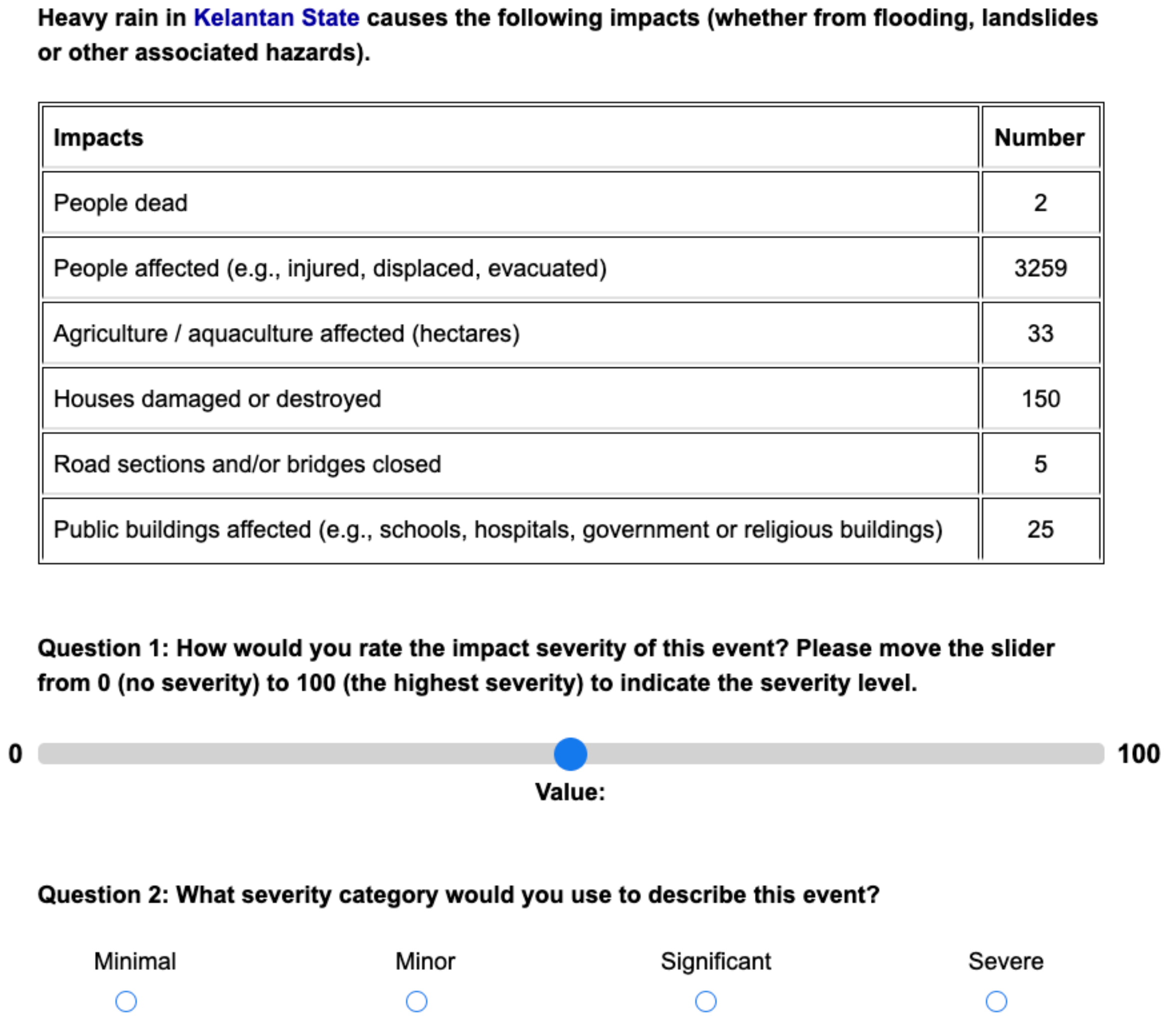
The two tasks in the study were designed to explore how weather scientists form severity judgments based on the six impacts (Judgment Analysis task), and to investigate their severity classifications for each impact (Impact Threshold Task). The Judgment Analysis task always preceded the Impact Threshold Task.

2.3.1 Judgment Analysis Task

Each participant viewed 70 hypothetical heavy rainfall events (consistent with the 10 to 1 ratio of events to cues prescribed for Judgment Analysis; Cooksey, 1996). Each event included a list of the impacts in a table for a country-specific location (see Figure 4). The order of the six impacts in the table was randomized between participants. Below the table, participants were asked to indicate the overall severity of the weather event in two formats. The first question asked “How would you rate the impact severity of this event?” with responses made on a scale from “0 (no severity)” to “100 (the highest severity)”. The starting point of the slider was 50 for every event. The second question asked participants to choose one of the four severity categories from the Risk Matrix (e.g., Minimal, Minor, Significant, Severe)[6] to describe the severity level of this event. We used a numerical severity judgment as well as the categorical severity classification to simplify the presentation of the results presented below (i.e., regression coefficients are much easier to understand for a continuous compared to a categorical outcome variable, when using an appropriate ordinal model for the latter).

Figure 4

An example event from the Judgment Analysis Task (Malaysia)



Heavy rain in Kelantan State causes the following impacts (whether from flooding, landslides or other associated hazards).

Impacts	**Number**
People dead	2
People affected (e.g., injured, displaced, evacuated)	3259
Agriculture / aquaculture affected (hectares)	33
Houses damaged or destroyed	150
Road sections and/or bridges closed	5
Public buildings affected (e.g., schools, hospitals, government or religious buildings)	25

Question 1: How would you rate the impact severity of this event? Please move the slider from 0 (no severity) to 100 (the highest severity) to indicate the severity level.

Question 2: What severity category would you use to describe this event?

○ Minimal ○ Minor ○ Significant ○ Severe

Note. The location of the heavy rainfall varies according to the country of the participants: Kelantan State in Malaysia; Jakarta in Indonesia; Metro Manila in Philippines; Hanoi in Vietnam.

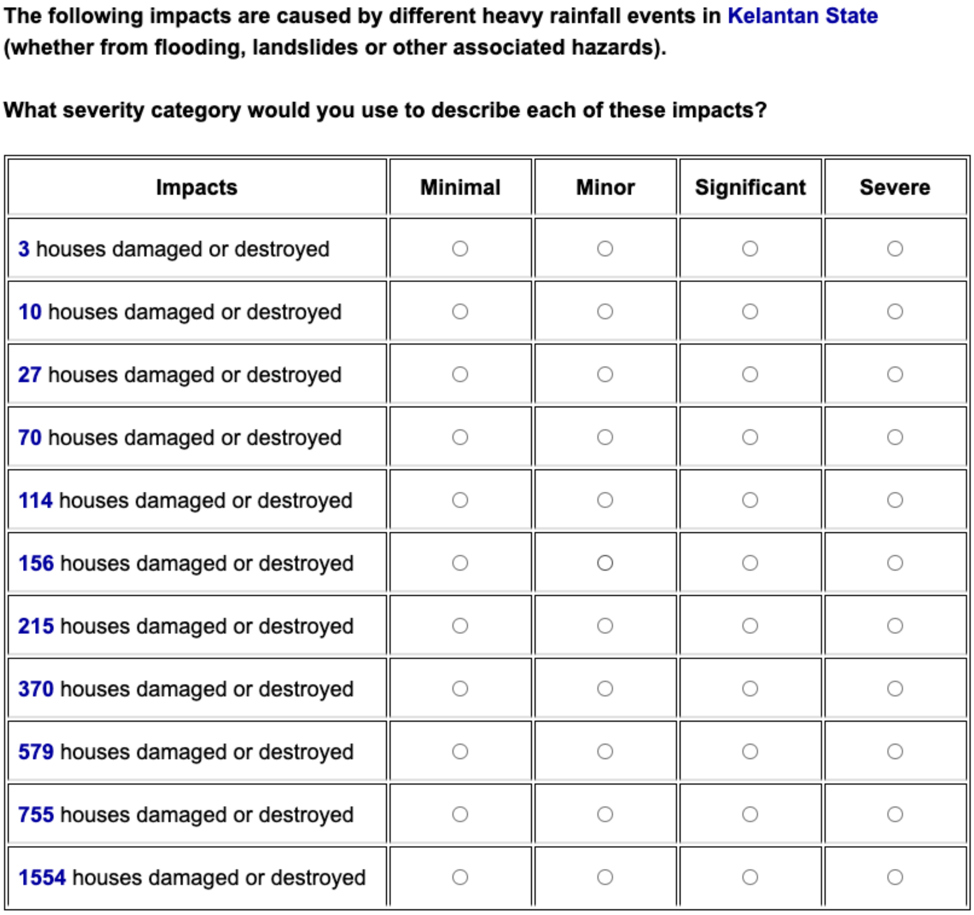
[6] To be consistent with the words used in partner countries, we used "Minimal, Minor, Significant and Severe" for Malaysia and Philippines, and they were translated into Bahasa for Indonesia. The terms "Minor, Potentially Dangerous, Dangerous and Very Dangerous" were used in Vietnam (translated into Vietnamese).

2.3.2 Impact Threshold Task

Participants saw six matrices on six pages, with each page corresponding to one of the six impact types (see Figure 5). For each matrix, they were told that "The following impacts are caused by different heavy rainfall events in Kelantan State[7] (whether from flooding, landslides or other associated hazards)". Then they were asked to select one severity classification (e.g., Minimal, Minor, Significant, Severe) to indicate the severity level of each of these impacts. The order of the six matrices was randomized between participants.

Figure 5

An example of the impact matrix of Houses damaged or destroyed (Malaysia)



The following impacts are caused by different heavy rainfall events in Kelantan State (whether from flooding, landslides or other associated hazards).

What severity category would you use to describe each of these impacts?

Impacts	Minimal	Minor	Significant	Severe
3 houses damaged or destroyed	○	○	○	○
10 houses damaged or destroyed	○	○	○	○
27 houses damaged or destroyed	○	○	○	○
70 houses damaged or destroyed	○	○	○	○
114 houses damaged or destroyed	○	○	○	○
156 houses damaged or destroyed	○	○	○	○
215 houses damaged or destroyed	○	○	○	○
370 houses damaged or destroyed	○	○	○	○
579 houses damaged or destroyed	○	○	○	○
755 houses damaged or destroyed	○	○	○	○
1554 houses damaged or destroyed	○	○	○	○

The numbers of the impacts in each matrix in the Impact Threshold Task were based on quantiles of the multivariate normal distribution estimated from real heavy rainfall events

[7] This is the location used in Malaysia's survey. Materials were customised for each country (see note in Figure 4).

(i.e., the same multivariate normal distribution used for creating the hypothetical heavy rainfall events). For the three impacts of “People affected (e.g., injured, displaced, evacuated)”, “Houses damaged and destroyed”, and “Agriculture/aquaculture affected”, we selected the 1% quantile as the first impact value. Then, for each participant anew, we randomly sampled one value from a uniform distribution with limits given by each pair of adjacent quantiles (1%, 10%, 20%, 30%, ... steps of 10 percentage points ..., 90%, 99%). This resulted in 11 judgments for each of these three impacts (see Appendix 3).

For the impacts of “People dead”, “Public buildings affected (e.g., schools, hospitals, government or religious buildings)” and “Road sections and/or bridges closed”, we used similar methods. Because of the even more extreme right skew of these impacts, however, we used fixed values of 1, 2 and 5 as their first three impact values. The impact of “Public buildings affected (e.g., schools, hospitals, government or religious buildings)” had another fixed value of 10 as the value of its fourth impact. Then we identified the 80%, 90%, 95% and 99% quantiles, and generated one random value from a uniform distribution with endpoints given by each pair of adjacent quantiles for each participant. Consequently, seven judgments were required for impact types “People dead” and “Road sections and/or bridges closed”, and eight judgments for “Public buildings affected (e.g., schools, hospitals, government or religious buildings)”.

2.4 Procedure

The study was developed using lab.js (Henninger et al., 2022) and delivered to participants using a JATOS server (https://www.jatos.org/). Each participant completed the study in a single session of approximately 30 minutes. After providing informed consent, participants were asked a series of demographic questions, such as the organisation that they work for, experience with impact-based weather forecasting, experience with the impact-based risk matrix, and whether they use IBF in daily life or not (see Table 1).

At the outset of the Judgment Analysis Task, participants received 5 practice trials. These trials were randomly selected from simulated events to provide participants with a general impression of the events and the mode of presentation. Responses to the practice trials were not analysed and participants were aware that these were practice trials. Specifically, they were given the following instructions:

> "Before the start of survey, you will now see 5 practice events, so that you can get a feel for the task before starting it for real.
>
> We will show you tables listing the impacts of heavy rain / flood events. Each table will be presented at the top of the screen, and you will be asked to rate the severity of the event both on a scale of 0-100, and also by indicating a severity level, which corresponds to the impact-based weather warning classification system: Minimal; Minor; Significant; Severe.
>
> While the events are simulated based on a series of real heavy rain events, you should treat the impacts as real.
>
> You are reminded that there are no right or wrong answers. It is really important that you answer these questions as YOU see appropriate as a weather forecaster / weather scientist."

After completing the practice trials, participants were told that they were now moving onto the main task: "On the following screens, you will be presented with 70 hypothetical heavy rain / flooding events, just like the ones in the practice. Each table will be presented at the top of the screen, and you will be asked to rate the severity level on two severity judgment questions listed below." They then completed the Judgment Analysis task (70 events).

After completing the Judgment Analysis Task, participants were given instructions for the Impact Threshold Task, and proceeded to complete it. The specific instructions were as follows:

> "Your answers to this task will enable us to undertake the appropriate analysis of your answers in Part 1.
>
> Each of the following 6 screens will present different values associated with a single impact. For each impact value, we ask you to rate whether that level of impact is Minimal, Minor, Significant or Severe.

> As before, there are no right or wrong answers. It is really important that you answer these questions as YOU see appropriate as a weather forecaster / weather scientist."

Finally, participants were thanked and debriefed.

3. Results

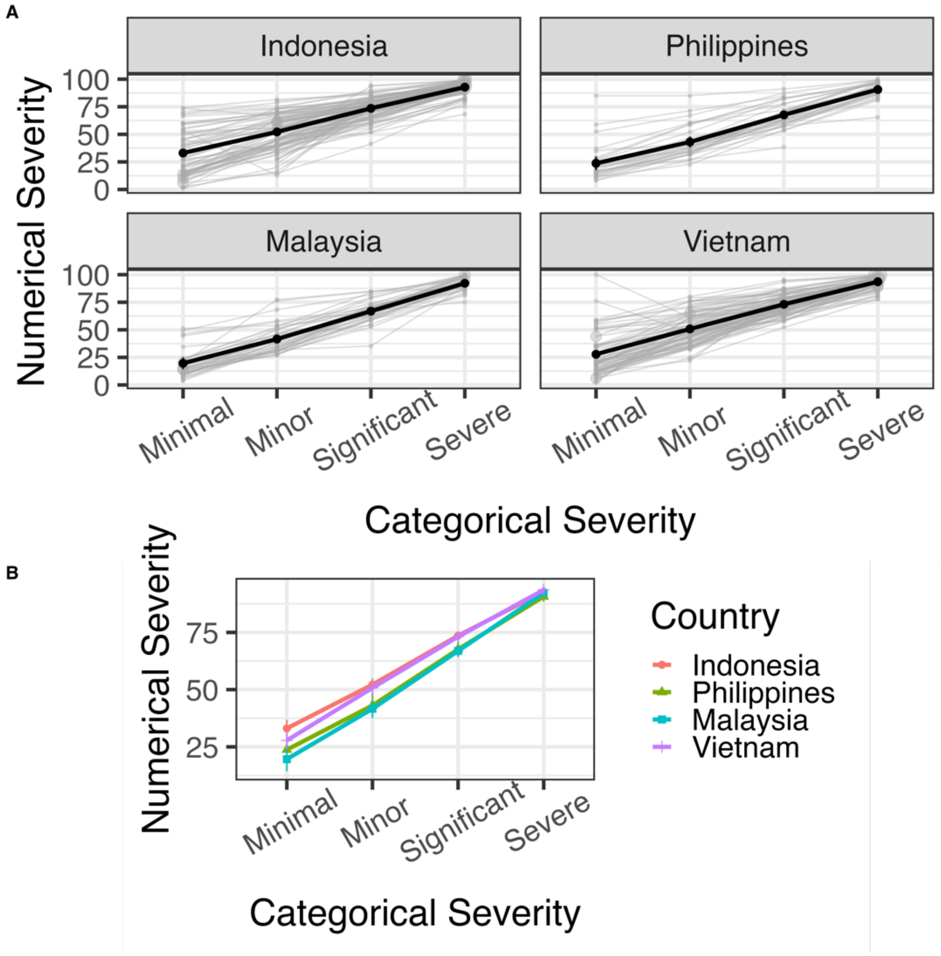
3.1 Correlations between numerical and categorical severity judgments

Participants provided numerical and categorical severity judgments. The former were included to considerably simplify the presentation of the results of the Judgement Analysis. However, a prerequisite for using the numerical severity judgements for the Judgement Analysis is a strong relationship between numerical and categorical ratings. Figure 6 shows that there is. Further evidence for this strong relationship comes from a linear mixed model ('Severity model') estimated with afex (an R software package for the Analysis of Factorial EXperiments, Singmann & Kellen, 2019)[8], with the numerical severity ratings as the dependent variable; the categorical severity ratings, Country, and their interaction as fixed effects; and by-participant random slopes for categorical severity ratings. In a second step, we then checked for the presence of a linear trend of categorical severity ratings on numerical severity ratings. In line with the visual impression of Figure 6, the linear trend was clearly significant (z = 52.86, p < .001). The linear trend was also significant in each country (Indonesia: z = 31.01, p < .001; Philippines: z = 21.90, p < .001; Malaysia: z = 26.41, p < .001; Vietnam: z = 33.93, p < .001).

[8] The Severity model specification in lme4 syntax (Bates et al., 2015) was as follows: Numerical severity judgments ~ Categorical severity judgments *Country + (Categorical severity judgments | id). The *p*-values were based on denominator degrees of freedom estimated with the Satterthwaite method (Kuznetsova et al., 2017). Estimating this model resulted in a solution with a gradient value slightly larger than the tolerance (0.004 > 0.002). Refitting the model without the correlations among random terms removed this issue and resulted in very similar parameter estimates and the same patterns of significance. In the text, we report the results from the maximal model. The following analyses on the maximal Impact models, including the maximal Impact model omitting zero deaths events, the maximal Position model and the maximal Interaction model, also showed the same convergence issues, and refitting models without the correlations among random terms removed these issues and revealed the similar patterns. We report the results of the maximal models in the text.

Figure 6

The relationship between numerical and categorical severity ratings.



Note. In Panel A, the grey lines correspond to mean ratings of individual weather scientists and the black lines correspond to the overall mean in those countries. Panel B only shows overall country means, with corresponding 95% confidence intervals.

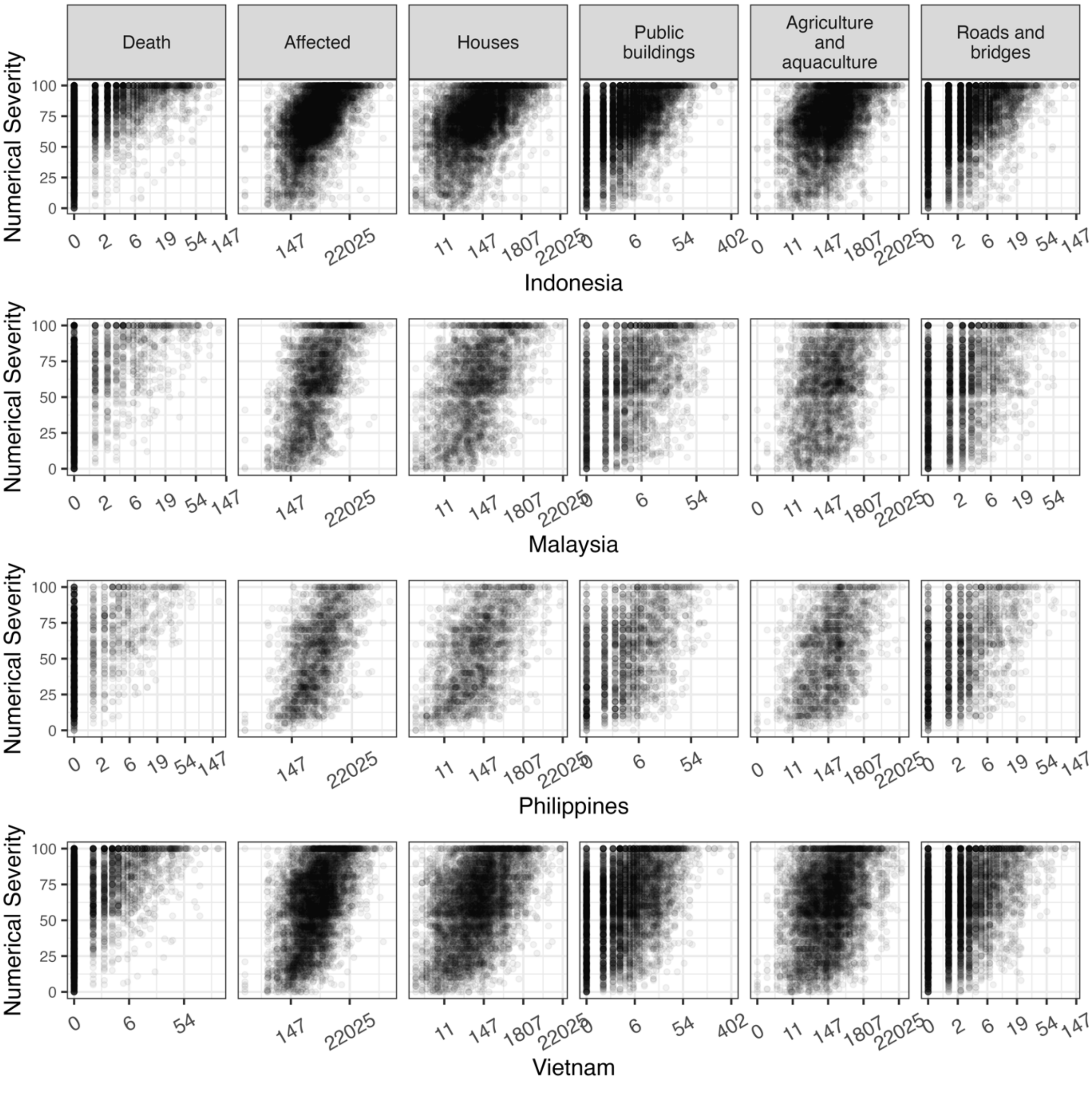
3.2 Correlations between impact values and numerical severity judgments

Figure 7 depicts the correlations between numerical severity ratings (from 1 to 100) and the provided values for each impact type. In this figure, and all reported analyses, the impact values were transformed using the log1P transformation which reduced the right-skew of all impacts and was the same data pre-processing as used for the real impacts from the impact database before analysis (see Section 2.3.1). Positive significant correlations were observed between severity ratings and all impact types (Death: *r* = 0.42; Affected: *r* = 0.51; Houses: *r* =

0.46; Public buildings: *r* = 0.39; Agriculture/Aquaculture: *r* = 0.36; Roads/Bridges: *r* =0.37). The higher the impact values, the higher the severity judgments.

Figure 7.

Relationship between impact values and numerical severity judgments in the four countries.



3.3 Judgment Analysis of weather scientists' severity judgments

To analyse participants' numerical severity ratings, we used a linear mixed effects model to utilise all the data, but also enable us to ascertain whether the influence of any impacts was moderated by Country. Specifically, the model (hereafter referred to as 'Impact model') included numerical severity ratings as the dependent variable, fixed effects for the six impacts and Country (entered as a factor with four levels), and the two-way interactions between each of the six impacts and Country. We used the maximal random effect structure justified by the design (Barr et al., 2013), with by-participant random intercepts, by-participant random slopes for the six impacts, as well as correlations among the random terms. To allow for a direct comparison of the magnitude of the relationship across impacts, the impact values were log1P transformed (see above) and *z*-standardized (i.e., for each impact variable, we first applied the log1P transformation, then we standardised by subtracting the mean and dividing by the standard deviation)[9].

Results of the Judgment Analysis are summarized in Table 2 and show significant main effects of all six impact types, such that higher values of each impact type resulted in higher numerical severity judgments ($p < .001$). This shows that all six impacts provided some unique contribution to participants' overall perceptions of the severity of a heavy rainfall event. Inspection of the standardised regression coefficients shows that increases in the Number of People Affected resulted in the biggest increase in severity judgments, followed by the Number of Deaths. There were interaction effects between Deaths and Country, between Affected and Country, and between Agriculture/Aquaculture and Country. Crucially, even though these results demonstrate some differences in the precise weightings of these impacts between countries, simple effects undertaken using *emmeans* (an R package for estimating marginal means and linear trends; Lenth, 2024), showed all impacts to be positive and significant in all countries (see Table A3 in Appendix 4).

Recall that only 30% of the heavy rainfall events provided to weather scientists included any deaths. A priori, we considered it possible that the inclusion of deaths would render all other impact types irrelevant for overall severity judgments. We therefore repeated the mixed model analysis with a data set from which we omitted all events with zero deaths. As in the Impact model on the full data, *all* impact types were found to be significant predictors of numerical severity judgments. In this model, Number of Deaths (F (1, 280.02) = 201.26, $p < .001$) and Number of People Affected (F (1, 257.79) = 80.15, $p < .001$) again were weighted

[9] The Impact model specification in lme4 syntax was as follows: Numerical severity judgments ~ (Deaths + Affected + Houses + Public buildings + Agriculture/Aquaculture + Roads/Bridges)* Country + (Deaths + Affected + Houses + Public buildings + Agriculture/Aquaculture + Roads/Bridges | id).

most heavily in severity judgments, followed by Houses (*F* (1, 309.25) = 25.87, *p* < .001), Agriculture/Aquaculture (*F* (1, 563.04) = 48.50, *p* < .001) and Roads/Bridges (*F* (1, 272.90) = 12.18, *p* < .001). Public buildings (*F* (1, 307.91) = 6.01, *p* = 0.01) received the least weight in their numerical severity.

While all six impact types improve the predictiveness of the overall model, it is worth drawing attention to the considerable interpersonal variation in the weightings of each impact type (as indicated by the standard deviations of the random slopes [Table 2, final column]). For example, the smallest SD is 1.26 for public buildings which has a fixed-effect coefficient of $\beta = 0.94$. Assuming the assumption of normally distributed individual-level effects holds reasonably well, this means that more than 1/6 of participants have a coefficient that is more than double the value of the fixed effect and more than 1/6 of participants even have a negative coefficient.[10] We will return to this issue in the Discussion, highlighting it as an important focus for future research.

Table 2

Coefficient estimates from Impact model with impacts as predictors, predicting heavy rainfall event severity ratings

Predictor	**df**	**F**	**p**	**Estimate**	**SE**	**Random effect (SD)**
(Intercept)	1, 273.72	3382.94	**<.001**	60.94	1.05	15.50
Deaths	1, 273.68	275.81	**<.001**	4.67	0.28	3.75
Affected	1, 270.64	353.19	**<.001**	7.25	0.39	4.92
Houses	1, 268.43	77.20	**<.001**	2.18	0.25	2.43
Public buildings	1, 265.21	34.95	**<.001**	0.94	0.16	1.26
Agriculture/ Aquaculture	1, 273.82	229.02	**<.001**	2.32	0.15	1.45
Roads/ Bridges	1, 272.66	63.99	**<.001**	1.36	0.17	1.61
Country	3, 273.72	13.90	**<.001**			
Death: Country	3, 273.32	7.39	**<.001**			
Affected: Country	3, 270.29	3.21	**0.024**			
Houses: Country	3, 267.00	1.01	0.388			
Public buildings: Country	3, 262.66	0.52	0.669			
Agriculture/ Aquaculture: Country	3, 273.81	5.66	**<.001**			
Roads/ Bridges: Country	3, 270.03	1.57	0.198			

[10] For a normal distribution, around 1/6 of data is larger than mean + 1*SD (0.94 + 1.26) and around 1/6 of data is smaller than mean – 1*SD (0.94 – 1.26).

Note. Significant predictors ($p < 0.05$) are shown in bold.

The previous analyses demonstrate that all six impact types are used to form severity judgments across participants. These results could stem from all participants generally using all six impacts, or subsets of participants relying on different subsets of impacts. Were the latter true, one would expect to see negative correlations between the by-participant random slopes of different impacts, indicating that participants who weight certain impacts strongly, weight other impacts less strongly (or not at all). The random slopes portrayed in Table 3 do not support such a conclusion. The preponderance of positive correlations is more supportive of the conclusion that weather scientists typically used all six impacts when making severity judgments in this task. As a further test, we investigated whether participants focussed on different impacts according to the order in which they were presented (as order was randomised between participants). Specifically, we categorised predictors according to their position in the table for any individual participant ('First', 'Second', 'Third', 'Fourth', 'Fifth', and 'Sixth'). We then ran a further linear mixed model with the numerical severity ratings as a dependent variable, the six *positions* of the six impacts, Country, and the interaction between each of the *Position* and Country as the fixed effects. As before, we employed the maximal random effect structure with by-participant random intercept and random slope for the six *positions* as well as the correlations among by-participant random terms ('Position model')[11]. The results of the maximal model are summarized in Table 4.

The Position model revealed two interesting results: 1) The first impact had the largest effect, with the size of the effects decreasing monotonically. This suggests that participants did generally weight the first impact most heavily; 2) More importantly, the model showed significant main effects of all six predictors ($p < 0.001$), implying that all variables in the event impact table were significant predictors for numerical severity judgment no matter which position they were in the impact table. This further supports our conclusion that all six impacts were considered in overall severity judgments.

[11] The Position model specification in lme4 syntax was as follows: Numerical severity judgments ~ (First + Second + Third + Fourth + Fifth + Sixth) * Country + (First + Second + Third + Fourth + Fifth + Sixth | id).

Table 3

Correlations between random slopes of impact types from the full data Impact model

	Deaths	**Affected**	**Houses**	**Public buildings**	**Agriculture/ Aquaculture**
Affected	-0.22				
Houses	-0.04	0.00			
Public buildings	0.22	-0.08	0.39		
Agriculture/ Aquaculture	0.20	0.22	0.04	0.42	
Roads/ Bridges	0.21	0.10	0.03	0.22	0.25

Table 4

Coefficient estimates from the Position model with impact Positions as predictors of severity ratings

Predictor	**df**	**F**	**p**	**Estimate**	**SE**	**Random effect (SD)**
First	1, 252.04	152.20	**<.001**	4.06	0.33	4.39
Second	1, 238.99	156.09	**<.001**	2.87	0.23	2.63
Third	1, 270.80	153.39	**<.001**	3.26	0.26	3.24
Fourth	1, 246.32	130.39	**<.001**	3.16	0.28	3.54
Fifth	1, 236.58	139.96	**<.001**	2.89	0.24	2.97
Sixth	1, 252.31	107.53	**<.001**	2.63	0.25	3.12
Country	3, 274.04	13.96	**<.001**			
First: Country	3, 252.41	0.24	0.87			
Second: Country	3, 237.02	1.09	0.36			
Third: Country	3, 268.94	2.70	**0.046**			
Fourth: Country	3, 246.18	1.00	0.40			
Fifth: Country	3, 237.49	1.30	0.28			
Sixth: Country	3, 250.34	0.90	0.44			

Note. We only report the estimated coefficients (beta) and SEs, for model terms corresponding to one coefficient (i.e., not for terms involving country). Significant predictors ($p < 0.05$) were shown in bold.

To investigate the influence of Experience with IBFs (see Table 1) on judgment policies (how overall severity judgments were derived from the impact information), we added a fixed effect of Experience into the Impact model, as well as 2-way interactions with each of the impacts[12]. The four response levels of Experience with IBFs were classified into two levels, in order to guarantee having enough data in each group. Hence, the responses with 'Little or no experience' and 'I have received training on it' were classified into the 'Inexperienced group' (N = 160), and the responses with 'Some experience' and 'A lot of experience' were classified into the 'Experienced group' (N = 118). We observed no significant main effects or interactions involving Experience ($p > 0.05$), suggesting the identified judgment policies were not moderated by Experience.

3.4 How do weather scientists combine severity levels of separate impact types into a single severity judgment?

The Judgment Analysis results are consistent with the use of a compensatory, weighted additive strategy, whereby all impact types are linearly combined in arriving at an overall severity judgment. In this section, we perform a further test of this idea by directly comparing compensatory and non-compensatory decision strategies using a descriptive data visualisation approach. We specifically test whether participants rely on a subset of impact types (a non-compensatory strategy), or use all impact types (a compensatory strategy). If, for example, five impact types are of a minimal severity level, but one is of a severe level, do weather scientists base their overall classifications on the maximum level ('severe'), the modal level ('minimal'), or some aggregation across the events (i.e., a compensatory strategy)? To perform this test, it was necessary to determine the thresholds for severity classifications for each individual impact type.

3.4.1 Severity classification thresholds for each individual impact type

In the Impact Threshold Task, participants provided categorical severity classifications for specified numerical values of individual impact types. We used these responses to estimate severity thresholds for each impact. For this, we estimated six ordinal regression models, one for each impact. More specifically, we estimated cumulative models with probit link

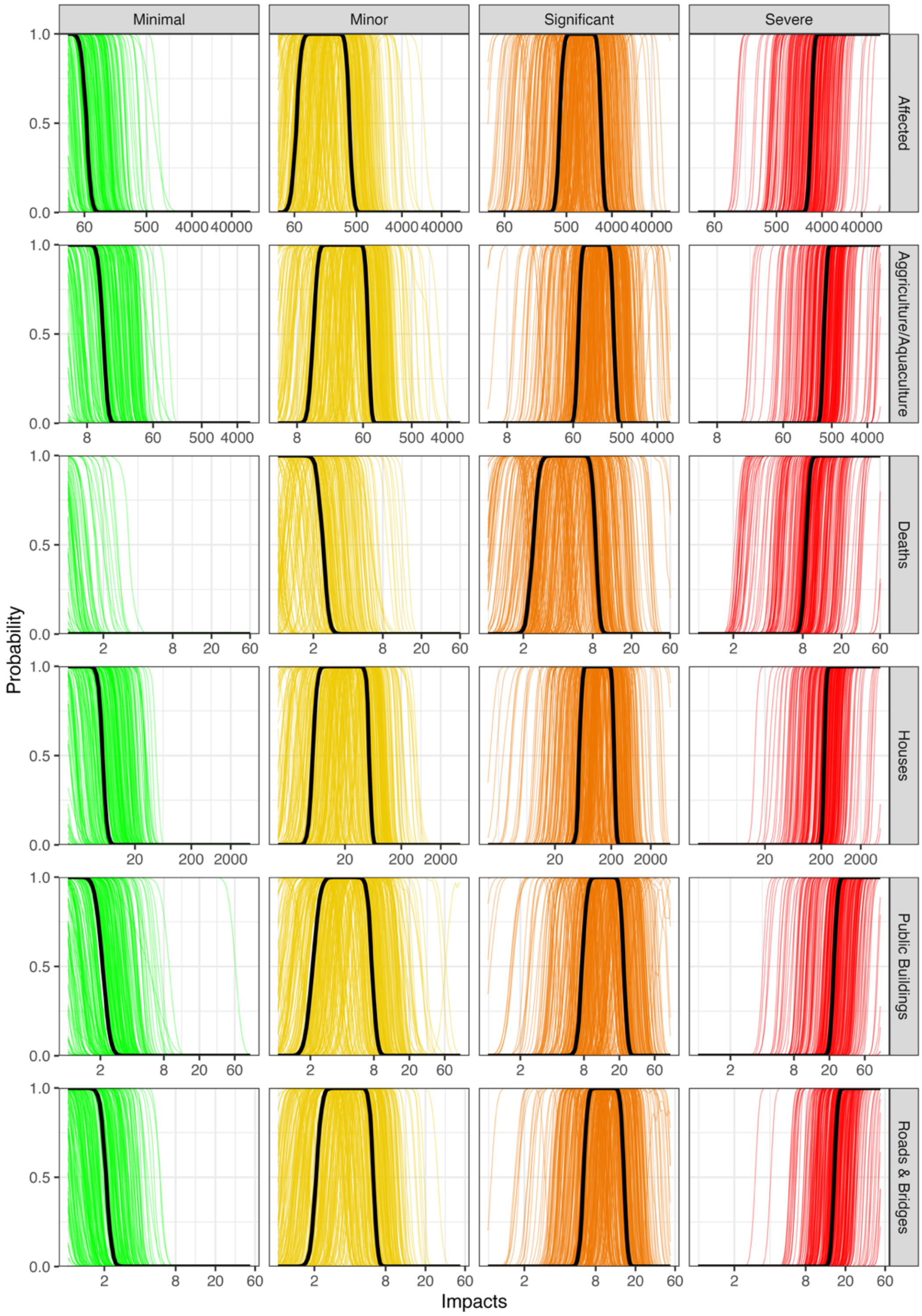
[12] The Impact model with the variable of Experience in lme4 syntax was as follows: Numerical severity judgments ~ (Deaths + Affected + Houses + Public buildings + Agriculture/Aquaculture + Roads/Bridges) * Country * Experience + (Deaths + Affected + Houses + Public buildings + Agriculture/Aquaculture + Roads/Bridges | id).

(Bürkner & Vuorre, 2019), with the categorical severity ratings as a dependent variable and the numerical impact values, as well as Country, as fixed-effect predictors. We also estimated by-participant random thresholds as well as a by-participant random slope for each impact value[13]. This by-participant random effect structure meant that the thresholds and resulting predictions were specific to each individual weather scientist, which resulted in models providing a good account of the observed data (see Appendix 6). The specificity of these thresholds means, for example, that the same number of people affected might be classified as Severe for Forecaster 1, but as Significant for Forecaster 2. When comparing the cumulative model with by-participant random terms (i.e., with idiosyncratic thresholds and idiosyncratic effects of impacts) with a cumulative model without by-participant random terms we saw clear qualitative and quantitative differences. The estimates of the fixed-effect, as well as the predicted categorical responses, differed dramatically between models. In the models with by-participant random terms, the predicted responses categories were much more certain (i.e., predicted probabilities near 1) compared to the model without the by-participant random terms where they were much more uncertain (i.e., predicted probabilities near 0.5). This result is in line with our finding of substantial random effect variances in the Judgement Analysis and suggests that inter-individual variance must be accounted for in the cumulative models. In other words, different weather scientists have markedly different perceptions of what constitutes a significant versus severe number of people affected (for example). This variance can be seen in Figure 8, which shows thresholds for each individual weather scientist (coloured lines), as well as overall mean thresholds (bold black lines).

[13] For example, the cumulative model of Death in brms syntax (Bürkner, 2017): Death severity classification ~ 1 + log1P (Death values)* Country + (cs(1)|id) + (0 + log1P(Death values)|id).

Figure 8

The fitted thresholds for the six impact types



Note. The coloured lines correspond to individual weather scientists' thresholds and the black lines correspond to the mean thresholds.

3.4.2. What decision strategy do weather scientists use to combine multiple impacts into an overall severity classification?

We used the unique individual thresholds for each participant to predict each participant's categorical severity classification for each impact in each impact table in the Judgment Analysis Task (i.e., for each numerical impact as exemplified in Figure 4). We then computed predictions of four possible decision strategies for each hypothetical heavy rainfall event. We used one compensatory strategy (the Mean strategy), and three non-compensatory strategies: Max, Mode (largest) and Mode (average), and compared the resulting predictions with participants' ('observed') severity classifications.

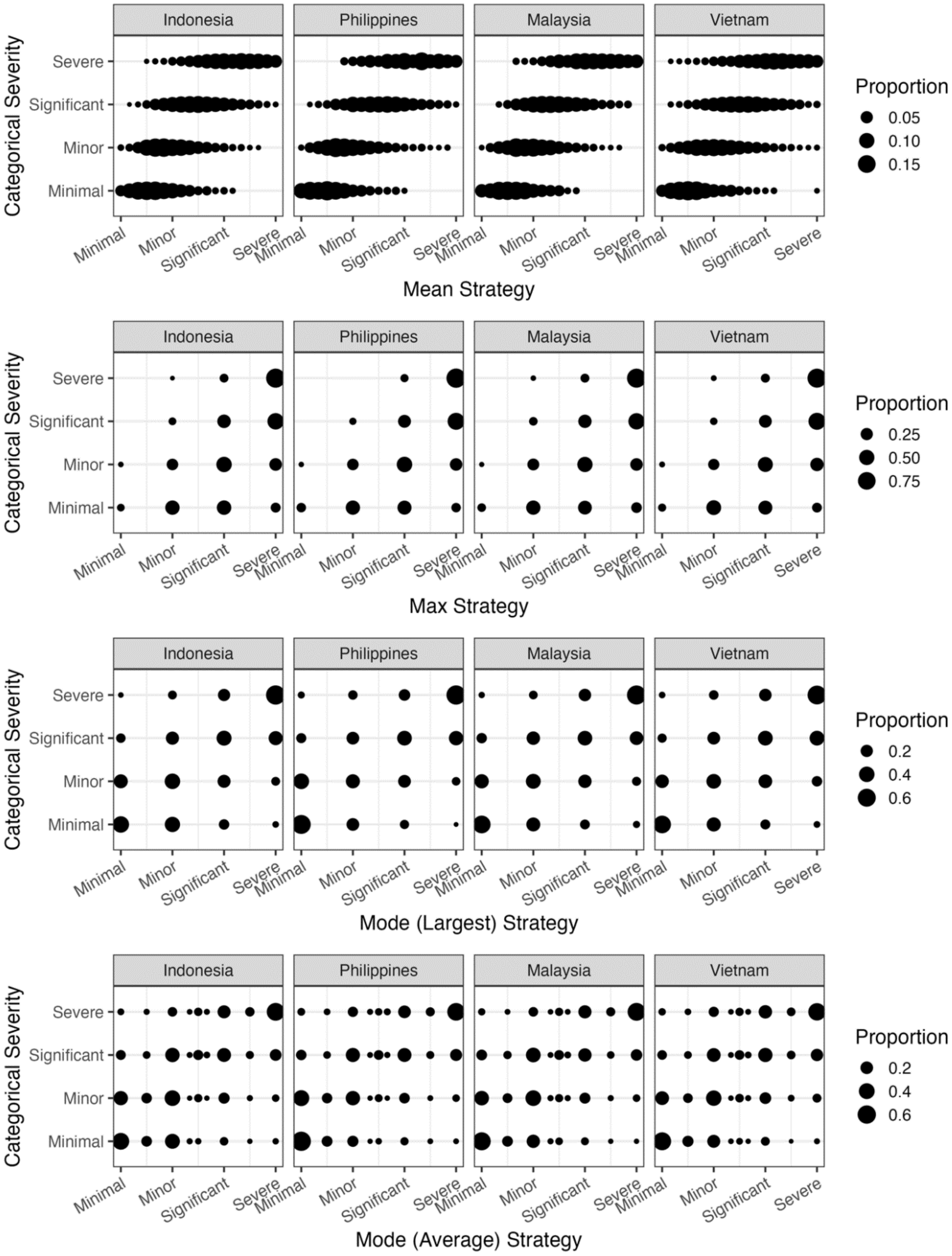
In order to compute the predictions of the Mean strategy, the four categorical severity levels were coded as follows: Minimal = 1, Minor = 2, Significant = 3, Severe = 4. The predicted severity classification is the mean of the numbers. For the Max strategy, the prediction is given by the maximum severity level. For the Mode strategy, the prediction is given by the modal severity level. In case of multiple modes (which occurred frequently), we used two different approaches[14]: (1) The Mode (largest) Strategy is the largest mode (the highest severity level); (2) The Mode (average) Strategy implied that the mean severity level of all modes was used to present the overall severity level of the event.

Figure 9 provides an overview of the relationship between the predictions derived from the four strategies described in the previous paragraph (on the *x*-axes) and the observed severity classifications (on the *y*-axes). For all of the shown strategies, we can see a clear positive relationship between the prediction made by the strategy and the weather scientists' actual severity classifications. In other words, all strategies seem to be able to account for the observed severity classifications to a similar degree based on the data shown in Figure 8. The only strategy that seems to make some clear qualitative mis-predictions was the Max strategy where the predicted classifications were somewhat higher than the observed classifications (the larger points are to the right of the main diagonal).

[14] For example, in a heavy rainfall event, two of the six impacts were rated as "Minimal", two of them were rated as "Significant" and the last two were rated as "Severe". Thus there were three modes in the event.

Figure 9

The use of different strategies in four countries



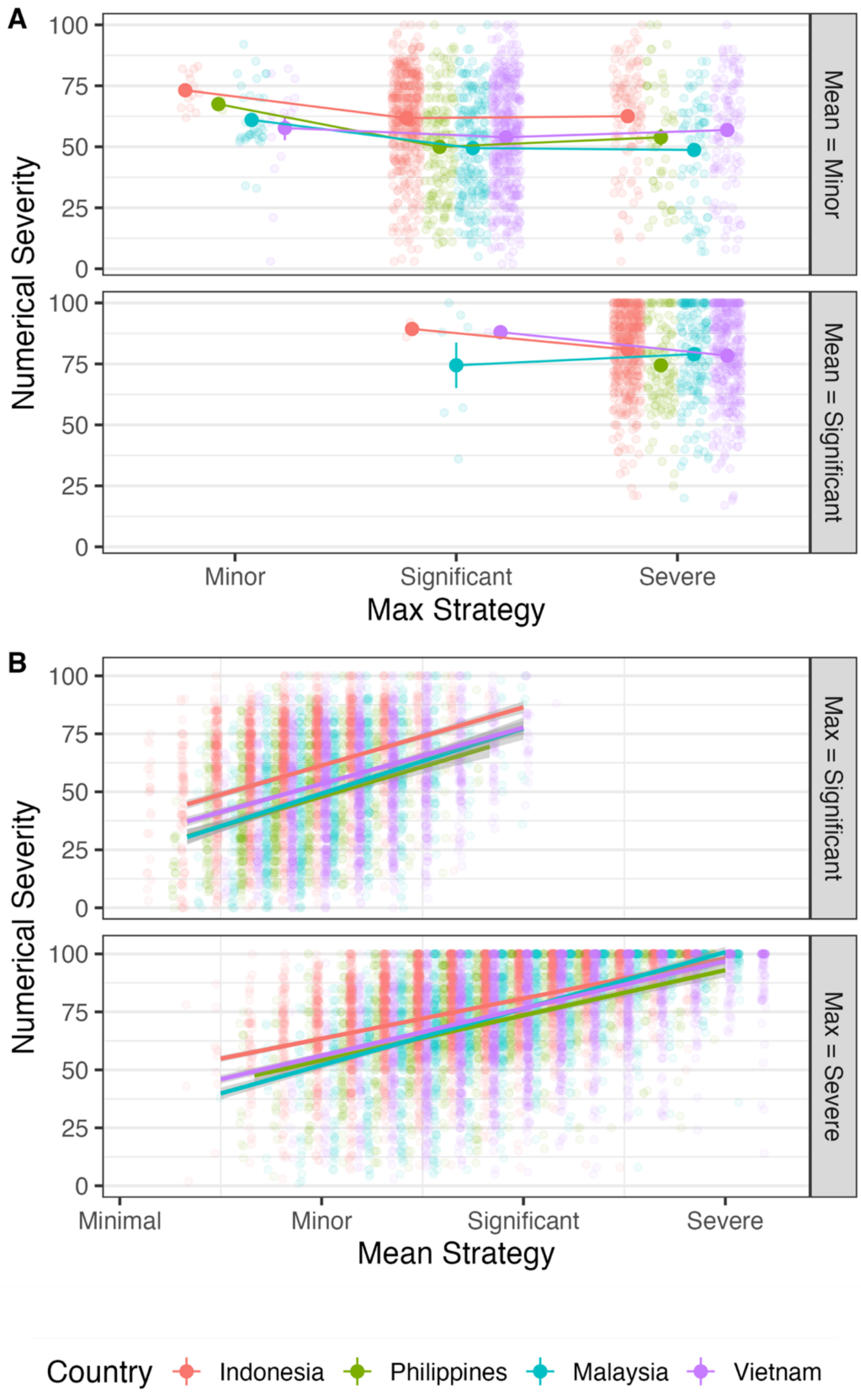
Note. The x-axes show the severity classification predicted by the strategy, whilst the y-axes show the observed severity classification. All trials (Judgment Analysis Task) are included. The size of the data points indicates the relative proportion of trials with that value within a given panel. Because both the Mean and Mode (average) strategy involve averaging the numerical category codes, the predicted strategy classifications can fall between categories.

Figure 9 suggests that the correlation among the different strategies is substantial and therefore looking at the overall predictions of the different strategies does not allow us to identify a single strategy participants might have used. In the following, we attempted to overcome this problem by directly comparing two strategies based on subsets of the data in which we control for the correlation between the compared strategies. More specifically, we always compare the compensatory mean strategy against one non-compensatory strategy, such that we hold the predictions of one of the two compared strategies constant and then evaluate whether a change in the other strategy is reflected with a change in the weather scientists' numerical severity rating. By doing this twice, once holding each strategy in a pair constant, visually it becomes readily apparent which of the two strategies has a stronger predictive ability. Note that for this purpose we use again the numerical severity ratings as the dependent variable as this provided for a clearer result (as discussed above, this is justifiable given the close relationship between the numerical ratings and the categorical classifications, see Figure 6).

The first such comparison is shown in Figure 10 which compares the Mean strategy with the Max strategy. In Panel A, the mean strategy is held constant at two levels, Minor and Significant (as these strategies represented the majority of responses). In Panel B, the Max Strategy is held constant at two levels, Significant and Severe (representing the majority of responses). Panel A shows that there was minimal influence of the Max Strategy when the Mean Strategy level remained constant. In contrast, when controlling for Max Strategy severity, the Mean Strategy levels tend to positively affect severity judgments (Panel B). It thus appears that participants' severity judgments are better predicted by the Mean strategy than the Max strategy. The same pattern was observed in comparison of the Mean Strategy against the Mode (largest) Strategy (Figure A5.1 in Appendix 5) and the Mode (average) Strategy (Figure A5.2 in Appendix 5). Overall, the results show that the (compensatory) Mean strategy appeared to be a better predictor of severity evaluations than other candidate strategies. Whilst additional analyses might show alternative compensatory strategies to be superior (e.g., a *Weighted Mean*), the aim of this analysis was to rule-out simpler non-compensatory strategies.

Figure 10

Comparison between Mean Strategy and Max Strategy



Note. In each panel, different colours represent different countries, with each datapoint reflecting one observation. Datapoints are plotted semi-transparently to avoid overplotting, such that darker points do indicate more data. In Panel A, individual data points are additionally jittered randomly on the x-axis. Furthermore, the large points represent means (with error bars representing the standard error of the mean). In Panel B, the solid lines show the regression lines.

3.4.3. Non-compensatory strategies using specific impact types.

The Judgment Analysis based on the Impact model found that Number of Deaths and Number of People Affected were the strongest predictors of overall severity judgments. Consequently, we wanted to compare the fully compensatory Mean strategy against strategies that focussed on only one of either of these two impacts. We did so again by focussing on the predicted impact severity classifications (i.e., analogous to Figure 10). As with the comparison against Max and Mode, a direct comparison of the Mean strategy with a strategy whereby weather scientists solely rely on Number of Deaths (Figure 11), or Number of People Affected (Figure 12), suggested the Mean strategy was a stronger predictor of the weather scientists' severity ratings. The Mean Strategy had an influence on the severity judgment after controlling for the Death Strategy and the Affected Strategy, but the influence was much weaker the other way round.

Figure 11

Comparison between Mean Strategy and Death Strategy

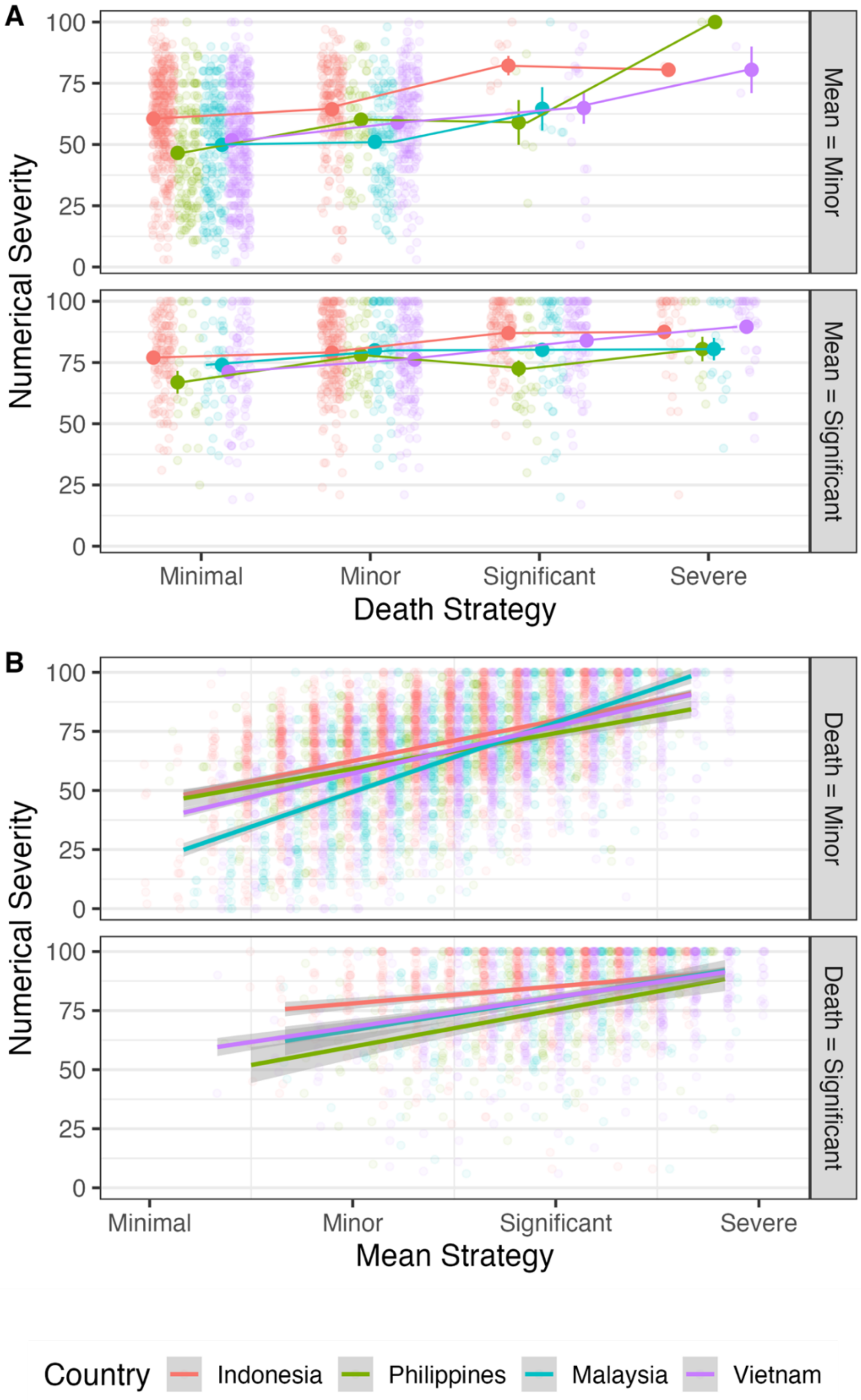
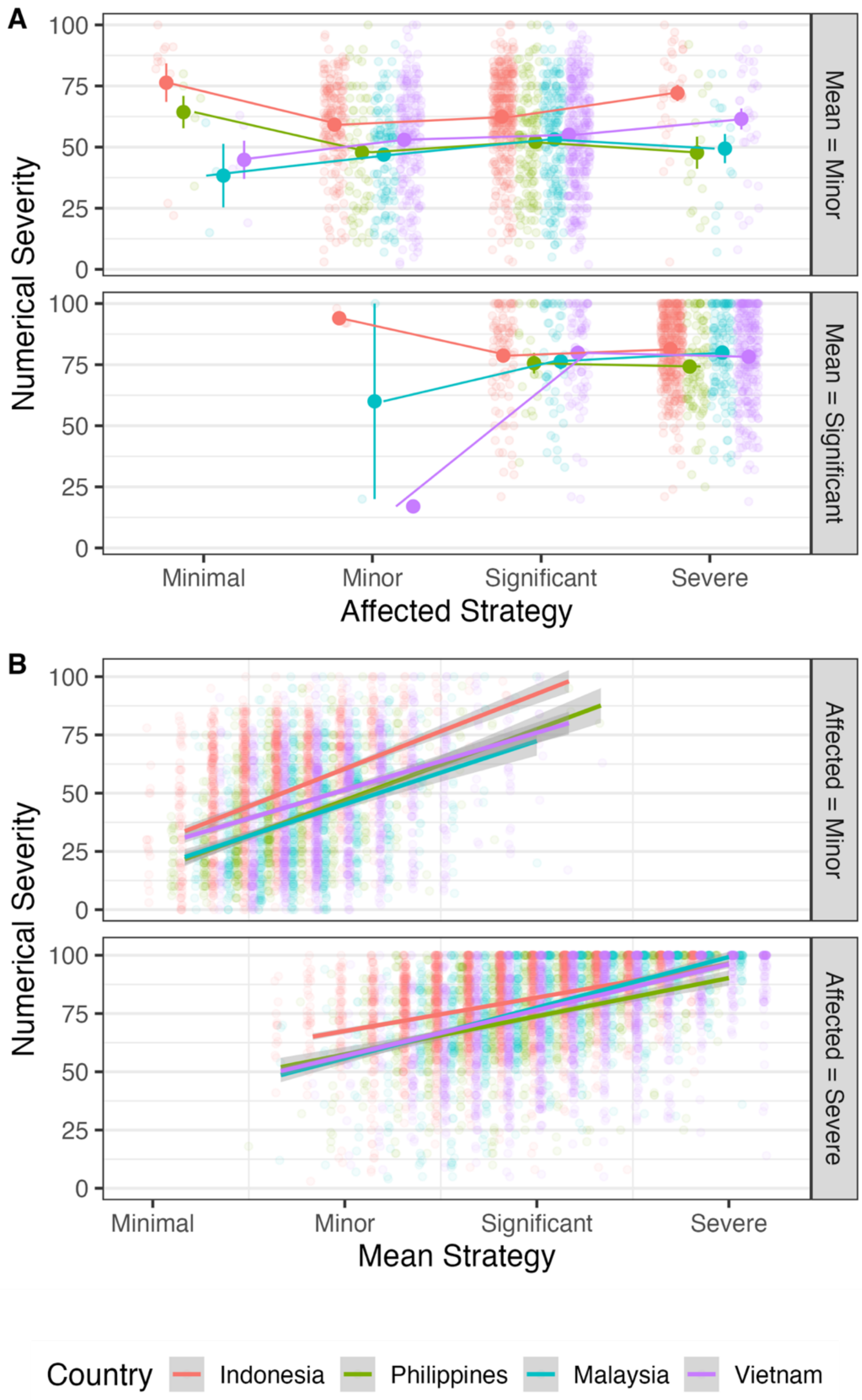


Figure 12

Comparison between Mean Strategy and Affected Strategy

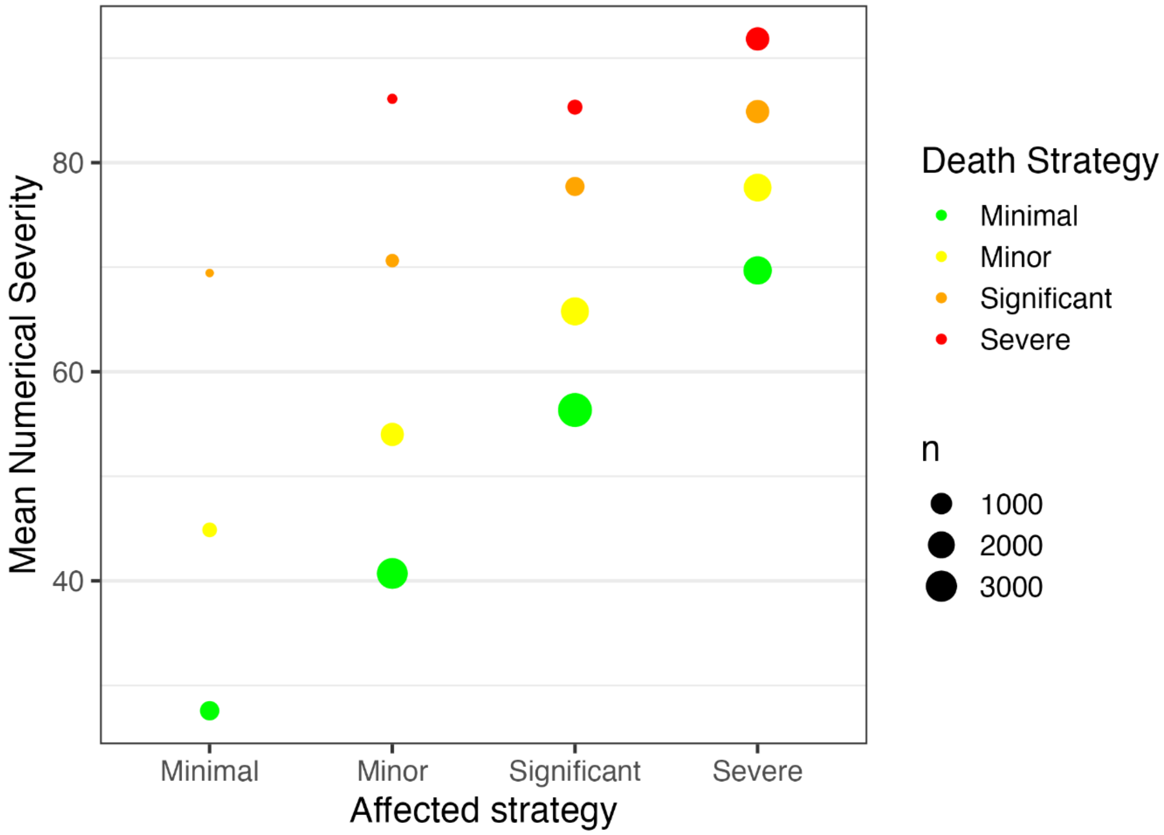


Thus far, we have paid no attention to possible interactions between individual predictors. This is because it is not possible to consider all possible interactions between the six impacts – the number of events required in the Judgment Analysis task would have been prohibitive. As an initial check that we are unlikely to be missing critical insights through not including interactions, we visualised the interaction between the Deaths strategy and the Affected strategy, as these were the most heavily weighted predictors in the Impact model. Figure 13 plots the mean numerical severity judgments for the combination of classifications predicted by the Death and Affected strategies. The main effects of the two variables are clear from Figure 13 - numerical severity judgments are higher as both Affected strategy predictions (on the x-axis) and Death strategy predictions (represented by colour) increase. Figure 13 also suggests an interaction, which is confirmed in a linear mixed effects model[15], *F* (1, 263.32) = 329.62, *p* <.001. When Deaths is Severe, the difference between the different severity levels of Affected is smaller than when Deaths is Minor. The most likely explanation of this interaction is a ceiling effect. When Deaths is already Significant or Severe, the severity cannot increase by much more as it is already approaching its maximum. We can also see that the two data points that largely drive the interaction (Death = Significant for Affected = Minimal; Death = Severe for Affected = Minor) represent only a minimal fraction of the overall data (each representing less than 50 observations, or 0.2% of all data). Because of this overall weak evidence for the importance of this interaction, we conclude that our overall conclusions are likely not compromised by excluding a consideration of interactions.

[15] The Interaction model specification was as follows: Numerical severity judgments ~ (Deaths * Affected) + (Deaths + Affected + Deaths * Affected | id).

Figure 13

The interaction of Death Strategy and Affected Strategy on numerical severity judgments



Note. Data points represents the mean numerical severity levels given by the combinations of predictions of Death strategy and predictions of Affected strategy. The size of data points reflects the number of observations. The interaction is illustrated by, for example, the greater effect of 'deaths' when the Affected strategy suggests Minimal severity than when it is Severe (the greater distance between the points in the leftmost 'column' of the figure versus the 'rightmost').

4. Discussion

In the current paper, we took a Judgment Analysis and decision strategy comparison approach to understand how weather scientists combine quantitative information about different impact types to form overall severity judgments. We found that weather scientists incorporated information from all six presented impact types in a compensatory process when forming an overall severity judgment for heavy rainfall events. Whilst all six impacts were found to be significant predictors for assessing severity levels, the two human factors ("Number of People Dead" and "Number of People Affected") received the most weight in weather scientists' judgments. We were able to rule out the possibility that this result was a group-level artefact of individual weather scientists each focussing on a different subset of impact types. This general result was observed in all four countries, notwithstanding slight variations in the precise weights assigned by weather scientists in the different countries. Whilst the data suggest a preponderance of a compensatory decision strategy in weather scientists, the precise weightings of impact types, and impact-value-to-severity-classification mappings (thresholds) varied considerably across individual weather scientists.

The compensatory strategy contrasts with the 'max' strategy used for aggregation in Aldridge et al. (2016). Meyer et al. (2007) argued that a disjunctive approach (such as the max strategy) was more appropriate than a more compensatory approach for spatial flood risk analyses. The disjunctive approach is a quick and simple rule that can help to screen the risk area since it requires only one threshold value is exceeded. In urgent situations, warnings need to be made in a timely manner, so a simpler and quicker strategy is preferred. It is possible that our weather scientists might use such a strategy where task characteristics alter the nature of the trade-off between time and accuracy (Huber & Kunz, 2007; Peng et al., 2019). A weighted mean has, however, been argued to be the optimal algorithm by which multiple cues should be combined (e.g., Cooksey, 1996; Frisch & Clemen, 1994). As accuracy concerns are likely more important in real forecasting scenarios, we maintain that these task characteristics are unlikely to encourage adoption of less optimal strategies. That said, given the large interpersonal differences in precise weightings and thresholds, it might be argued to be beneficial to include some consideration of error cost functions in the final, communicated, severity level, so that more costly errors (e.g., forecast recipients underestimating the impacts associated with a particular severity communication) can be avoided (Batchelor & Peel, 1998; Harris et al., 2009; Lawrence & O'Connor, 2005; Whiteley & Sahani, 2008; see also Liefgreen et al., 2024).

Our results also deviate from the majority of studies applying judgment analysis to expert judgment, where experts typically rely on a subset of cues available (c.f., Brehmer & Brehmer, 1988). As highlighted in the Introduction, however, such a result is by no means universal, with a sizeable minority of studies finding experts to aggregate across a large number of provided cues (see e.g., Chewning & Harrell, 1990; Taylor & Wilsted, 1974; Wilsted et al., 1975; White et al., 2018). In the current study, we only included a relatively small number of cues. The selection of these cues was based upon analysis of the impact types that are recorded (Wyatt et al., 2023), along with detailed focus discussions undertaken with experts sampled from the same population as the weather scientists who participated in the main study (many of those would have participated in the main study as well as the focus group). Thus, the observation that all six impacts contributed to overall severity judgments might suggest that the experts in the focus discussions had good insight into the relevant factors for determining the overall severity of a heavy rainfall event.

Future work, however, might seek to specifically test the question of the degree of insight weather scientists have into their judgment processes (see Cooksey, 1996). Such research should ensure that methods for assessing such awareness are suitably sensitive (see e.g., Lagnado et al., 2006; Newell & Shanks, 2014, 2023; Persaud et al., 2007).

4.1 Implications and future directions

This study investigated how weather scientists combine impact severity levels to form overall severity judgments for IBWs. One crucial observation was that there was considerable inter-forecaster variation in subjective thresholds of numerical impacts. This adds to previous observations (in experimental settings) of great variation in *probability* thresholds used by National Weather Service forecasters before issuing warnings (Trujillo-Falcón et al., 2022), and between broadcast meteorologists when deciding television coverage types to communicate impactful weather information to the public (Obermeier et al., 2022). Addressing the inter-individual differences between weather scientists is therefore a key research priority, further highlighted in the present work. These large individual differences prohibit the current findings from informing evaluations of IBF using quantitative information from database sources (see e.g., Wyatt et al., 2023). It was not possible to identify a common threshold for categorical severity classifications from numerical impact information. Whilst highlighting this variability is a good first step to addressing this issue, future work should seek to identify methods to improve consistency between weather scientists within individual countries. Finally, it may be desirable to develop a commonly recognized and standard guideline of severity threshold definitions for IBF within individual countries.

Although we showed that weather scientists' overall severity judgments were best predicted by a Mean strategy here, people with different roles and responsibilities to these scientists might use different strategies. Aldridge et al. (2016), for example, reported that disaster managers agreed that a Max strategy was most informative to enable quick computations and to avoid unexpected risks and consequences. Indicating that this is a desirable strategy is not, however, the same as using that strategy in classifications. Future work might seek to determine how stakeholders such as the disaster managers in Aldridge et al.'s work aggregate information about different impacts to inform overall severity judgments.

If IBWs are to be communicated to the general public, the public's understanding of these warnings is critical (e.g., Taylor et al., 2019, 2024). The examination of the general public's severity judgments is thus another important research direction. The public receive IBWs in many countries. Whether these are of a simplified nature (e.g., 'there is a medium likelihood of significant impacts from heavy rain'), or incorporate more detailed information about the hazard, source and impact (Casteel, 2016; Weyrich et al., 2018), it is still desirable that the public's perception of 'significant impacts' (and indeed 'medium likelihood') matches that intended by the weather scientists issuing the warnings (see also Potter et al., 2021). Only in this instance will the public be prepared for the forecasted weather impacts. Before engaging in such research, however, it is necessary to address the extant inter-individual differences between weather scientists.

4.2 Conclusion

Overall, weather scientists pay attention to all impact types, whilst giving more weight to human factors ('Number of People Dead' and 'Number of People Affected') when forming overall severity judgments. These results demonstrated a shared understanding of the importance of impacts between weather scientists from four countries in Southeast Asia. The consistency in the overall process was offset by the high levels of inter-forecaster variability in the precise weightings given to each impact type. As the utilization of IBF in Southeast Asia is still at a relatively early stage, helping weather scientists to form stable and consistent severity classifications will be beneficial for leveraging the full potential of IBF, as well as enabling subsequent evaluation of IBF.

Acknowledgments

This work and its contributors (Xiaoxiao Niu [UCL], Adam JL Harris [UCL]; Agie W Putra and Riefda Novikarany [BMKG]; Azlai Taat [MetMalaysia] and Sazali Osman [DID]; Lam Hoang and Diep Quang Tran [NCHMF]; Lorenzo A Moron and Jehan S Panti [PAGASA]; and Faye Wyatt and Rebecca Beckett [UK Met Office]) were funded by the Met Office Weather and Climate Science for Service Partnership (WCSSP) Southeast Asia project which is supported by the Department for Science, Innovation & Technology (DSIT). Our huge thanks go to the in-country partners and all of our participants – without this help, the work could not have happened.

References

AHA Center. (2023). ASEAN Weekly Disaster Update Week 2 (16 –22 January 2023). *Reliefweb*. Retrieved from https://reliefweb.int/report/indonesia/asean-weekly-disaster-update-week-2-16-22-january-2023

Aldridge, T., Gunawan, O., Moore, R. J., Cole, S. J., Boyce, G., & Cowling, R. (2020). Developing an impact library for forecasting surface water flood risk. *Journal of Flood Risk Management, 13*(3), e12641. https://doi.org/10.1111/jfr3.12641

Aldridge, T., Gunawan, O., Moore, R. J., Cole, S. J., & Price, D. (2016). *A surface water flooding impact library for flood risk assessment.* In E3S Web of *Conferences, 7, p. 18006*. EDP Sciences. https://doi.org/10.1051/e3sconf/20160718006

Baker, S., & Thompson, C. (2012). Initiating artificial nutrition support: a clinical judgement analysis. *Journal of human nutrition and dietetics*, *25*(5), 427-434. https://doi.org/10.1111/j.1365-277X.2012.01260.x

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68(3), 255–278.* https://doi.org/10.1016/j.jml.2012.11.001

Batchelor, R., & Peel, D. A. (1998). Rationality testing under asymmetric loss. *Economics Letters, 61*(1), 49-54. https://doi.org/10.1016/S0165-1765(98)00157-8

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67(1).* https://doi.org/10.18637/jss.v067.i01

Beckett, R., & Hartley, A. (2020). Progress on the development of impact based forecasting in South East Asia. *Met Office*, 1-87.

Boult, V. L., Black, E., Abdillahi, H. S., Bailey, M., Harris, C., Kilavi, M., ... & Todd, M. C. (2022). Towards drought impact-based forecasting in a multi-hazard context. *Climate Risk Management*, *35*, 100402. https://doi.org/10.1016/j.crm.2022.100402

Brehmer, A., & Brehmer, B. (1988). What have we learned about human judgment from thirty years of policy capturing?. In *Advances in psychology*, 54, 75-114. North-Holland. https://doi.org/10.1016/S0166-4115(08)62171-8

Browne, B. A. , & Gillis, J. S. . (1982). Evaluating the quality of instruction in art: a social judgment analysis. *Psychological Reports, 50*(3), 955-962. https://doi.org/10.2466/pr0.1982.50.3.955

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1-28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P. C., & Vuorre, M. (2019). Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science, 2*(1), 77-101. https://doi.org/10.1177/2515245918823199

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., , Goodrich, B., Betancourt, M., Brubaker, M. A., . . . Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software., 76*(1), 1-32. https://doi.org/10.18637/jss.v076.i01

Casteel, M. A. (2016). Communicating increased risk: An empirical investigation of the National Weather Service's impact-based warnings. *Weather, Climate, and Society*, *8*(3), 219-232. https://doi.org/10.1175/WCAS-D-15-0044.1

Centre for Excellence in Disaster Management and Humanitarian Assistance. (2019). Malaysia Disaster Management Reference Handbook. *Reliefweb.* Retrieved from https://reliefweb.int/report/malaysia/malaysia-disaster-management-reference-handbook-june-2019

Chau, V. N., Cassells, S., & Holland, J. (2015). Economic impact upon agricultural production from extreme flood events in Quang Nam, central Vietnam. *Natural Hazards, 75*(2), 1747-1765. https://doi.org/10.1007/s11069-014-1395-x

Chewning Jr, E. G., & Harrell, A. M. (1990). The effect of information load on decision makers' cue utilization levels and decision quality in a financial distress decision task. *Accounting, Organizations and Society*, *15*(6), 527-542. https://doi.org/10.1016/0361-3682(90)90033-Q

Cole, S. J., Moore, R. J., Wells, S. C., & Mattingley, P. S. (2016). *Real-time forecasts of flood hazard and impact: some UK experiences.* E3S Web of Conferences.

Cole, S. J., Moore, R. J., Wells, S. C., & Mattingley, P. S. (2016).*Real-time forecasts offlood hazard and impact: Some UK experiences*. In E3S Web of *Conferences, 7, p. 18015*. https://doi.org/10.1051/e3sconf/20160718015

Cooksey, R. W. (1996). *Judgment analysis: Theory, methods, and applications*: Academic press.

Dhami, M. K., & Harries, C. (2001). Fast and frugal versus regression models of human judgement. *Thinking & Reasoning, 7*(1), 5-27. https://doi.org/10.1080/13546780042000019

Dhami, M. K., Hertwig, R., & Hoffrage, U. (2004). The role of representative design in an ecological approach to cognition. *Psychological bulletin, 130*(6), 959-988. https://doi.org/10.1037/0033-2909.130.6.959

Doyle, E. E., McClure, J., Paton, D., & Johnston, D. M. (2014). Uncertainty and decision making: Volcanic crisis scenarios. *International Journal of Disaster Risk Reduction, 10*, 75-101. https://doi.org/10.1016/j.ijdrr.2014.07.006

ECHO. (2023). Madagascar - Tropical storm CHENESO, update (DG ECHO, UN OCHA, GDACS, MeteoFrance La Reunion, BNGRC, Meteo Madagascar) (ECHO Daily Flash of 20 January 2023). *Reliefweb*. Retrieved from https://reliefweb.int/report/madagascar/madagascar-tropical-storm-cheneso-update-dg-echo-un-ocha-gdacs-meteofrance-la-reunion-bngrc-meteo-madagascar-echo-daily-flash-20-january-2023

UN ESCAP, U. (2021). Manual for operationalizing impact-based forecasting and warning services (IBFWS). https://hdl.handle.net/20.500.12870/4544

Ettenson, R., Shanteau, J., & Krogstad, J. (1987). Expert judgment: Is more information better?. *Psychological reports*, *60*(1), 227-238. https://doi.org/10.2466/pr0.1987.60.1.227

Foerster, J. F. (1979). Mode choice decision process models: a comparison of compensatory and non-compensatory structures. *Transportation Research Part A: General, 13*(1), 17-28. https://doi.org/10.1016/0191-2607(79)90083-9

Frisch, D., & Clemen, R. T. (1994). Beyond expected utility: Rethinking behavioral decision research. *Psychological Bulletin, 116*(1), 46–54. https://doi.org/10.1037/0033-2909.116.1.46

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological Review, 103*(4), 650-669. https://doi.org/10.1037/0033-295X.103.4.650

Gigerenzer, G., Hertwig, R., & Pachur, T. (2011). *Heuristics: The foundations of adaptive behavior*: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199744282.001.0001

Gigerenzer, G., & Todd, P. M. (1999). Fast and frugal heuristics: The adaptive toolbox. In *Simple heuristics that make us smart* (pp. 3-34): Oxford University Press.

Global Facility for Disaster Reduction and Recovery. (2016). Country Profile: The Philippines*. GFDRR.* Retrieved from https://www.gfdrr.org/sites/default/files/publication/PHILIPPINES2016.pdf

Global Facility for Disaster Reduction and Recovery. (2018). Mainstreaming Disaster Resilience in Vietnam. *GFDRR.* Retrieved from https://www.unisdr.org/preventionweb/files/43495_11vietnam.pdf

Harries, C., St. T. Evans, J., Dennis, I., & Dean, J. (1996). A clinical judgement analysis of prescribing decisions in general practice. *Le Travail Humain*, 87-109. https://www.jstor.org/stable/40659992

Harris, A. J., Corner, A., & Hahn, U. (2009). Estimating the probability of negative events. *Cognition, 110*(1), 51-64. https://doi.org/10.1016/j.cognition.2008.10.006

Harrison, S. E., Potter, S. H., Prasanna, R., Doyle, E. E., & Johnston, D. (2021). 'Where oh where is the data?': Identifying data sources for hydrometeorological impact forecasts and warnings in Aotearoa New Zealand. *International Journal of Disaster Risk Reduction*, *66*, 102619. https://doi.org/10.1016/j.ijdrr.2021.102619

Haryanto, B., Lestari, F., & Nurlambang, T. (2020). Extreme events, disasters, and health impacts in Indonesia. *Extreme Weather Events and Human Health: International Case Studies*, 227-245. https://doi.org/10.1007/978-3-030-23773-8_16

Hemingway, R., & Robbins, J. (2020). Developing a hazard-impact model to support impact-based forecasts and warnings: The Vehicle OverTurning (VOT) Model. *Meteorological Applications, 27*(1), e1819. https://doi.org/10.1002/met.1819

Henninger, F., Shevchenko, Y., Mertens, U. K., Kieslich, P. J., & Hilbig, B. E. (2022). lab.js: A free, open, online study builder. *Behavior Research Methods, 54(2), 556–573.* https://doi.org/10.3758/s13428-019-01283-5

Huber, O., & Kunz, U. (2007). Time pressure in risky decision-making: effect on risk defusing. *Psychology Science, 49*(4), 415-426. http://www.communicationcache.com/uploads/1/0/8/8/10887248/time_pressure_in_risky_decision-making-_effect_on_risk_defusing.pdf

Hurlbert, S. H. (1984). Pseudoreplication and the design of ecological field experiments. *Ecological monographs, 54*(2), 187-211. https://doi.org/10.2307/1942661

Jenkins, S. C., Harris, A. J., & Lark, R. M. (2019). When unlikely outcomes occur: the role of communication format in maintaining communicator credibility. *Journal of risk research, 22*(5), 537-554. https://doi.org/10.1080/13669877.2018.1440415

Jenkins, S. C., Putra, A. W., Ayuliana, S., Novikarany, R., Khalid, N. M., Mamat, C. S. N. C., ... & Harris, A. J. (2022). Investigating the decision thresholds for impact-based warnings in South East Asia. *International Journal of Disaster Risk Reduction*, *76*, 103021. https://doi.org/10.1016/j.ijdrr.2022.103021

Kaltenberger, R., Schaffhauser, A., & Staudinger, M. (2020). "What the weather will do"– results of a survey on impact-oriented and impact-based warnings in European NMHSs. *Advances in Science and Research*, *17*, 29-38. https://doi.org/10.5194/asr-17-29-2020

Kuo, Y. Y., & Liang, K. Y. (2004). Human judgments in New York state sales and use tax forecasting. *Journal of Forecasting*, *23*(4), 297-314. https://doi.org/10.1002/for.914

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software, 82, 1–26.* https://doi.org/10.18637/jss.v082.i13

Lagnado, D. A., Newell, B. R., Kahan, S., & Shanks, D. R. (2006). Insight and strategy in multiple-cue learning. *Journal of Experimental Psychology: General*, *135*(2), 162. https://doi.org/10.1037/0096-3445.135.2.162

Lawrence, M., & O'Connor, M. (2005). Judgmental forecasting in the presence of loss functions. *International Journal of Forecasting, 21*(1), 3-14. https://doi.org/10.1016/j.ijforecast.2004.02.003

Lenth, R. (2024). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.10.3, <https://CRAN.R-project.org/package=emmeans>.

Liefgreen, A., Jenkins, S. C., Osman, S., Moron, L. A., Monteverde, M. C. A., Cayanan, E. O., ... & Harris, A. J. (2024). Severity influences categorical likelihood communications: A case study with Southeast Asian weather forecasters. *Scientific Reports*, *14*(1), 14607. https://doi.org/10.1038/s41598-024-64399-5

Malczewski, J. (1999). *GIS and multicriteria decision analysis*: John Wiley & Sons.

Malczewski, J., & Rinner, C. (2015). *Multicriteria decision analysis in geographic information science* (Vol. 1): Springer.

Met Office. (2017). What are the National Severe Weather Warning Service Impact tables? Retrieved from https://www.metoffice.gov.uk/weather/guides/severe-weather-advice

Met Office. (2023). National Severe Weather Warning Service (NSWWS). Retrieved from https://www.metoffice.gov.uk/binaries/content/assets/metofficegovuk/pdf/data/pwms_nswws.pdf

Meyer, V., Haase, D., & Scheuer, S. (2007). GIS-based multicriteria analysis as decision support in flood risk management. Retrieved from https://www.econstor.eu/handle/10419/45237

Mitheu, F., Stephens, E., Petty, C., Ficchì, A., Tarnavsky, E., & Cornforth, R. (2023). Impact-based flood early warning for rural livelihoods in Uganda. *Weather, climate, and society*, *15*(3), 525-539. https://doi.org/10.1175/WCAS-D-22-0089.1

Moore, R. J., Cole, S. J., Dunn, S., Ghimire, S., Golding, B. W., Pierce, C. E., . . . Speight, L. (2015). Surface water flood forecasting for urban communities. CREW Report CRW2012_03, 32pp. https://nora.nerc.ac.uk/id/eprint/510114

Murphy, A. H., & Daan, H. (1984). Impacts of feedback and experience on the quality of subjective probability forecasts. comparison of results from the first and second years of the zierikzee experiment. *Monthly Weather Review, 112*(3), 413-423. https://doi.org/10.1175/1520-0493(1984)112<0413:IOFAEO>2.0.CO;2

Murphy, A. H., & Winkler, R. L. (1974a). Credible interval temperature forecasting: some experimental results. *Monthly Weather Review, 102*(11), 784-794. https://doi.org/10.1175/1520-0493(1974)102<0784:CITFSE>2.0.CO;2

Murphy, A. H., & Winkler, R. L. (1974b). Probability forecasts: A survey of National Weather Service forecasters. *Bulletin of the American Meteorological Society, 55*(12), 1449-1453. https://doi.org/10.1175/1520-0477(1974)055<1449:PFASON>2.0.CO;2

Murphy, A. H., & Winkler, R. L. (1984). Probability forecasting in meteorology. *Journal of the American statistical Association, 79*(387), 489-500. https://doi.org/10.1080/01621459.1984.10478075

Newell, B. R., & Shanks, D. R. (2014). Unconscious influences on decision making: A critical review. *Behavioral and brain sciences*, *37*(1), 1-19. https://doi.org/10.1017/S0140525X12003214

Newell, B. R., & Shanks, D. R. (2023). *Open minded: Searching for truth about the unconscious mind*. MIT Press.

Nkiaka, E., Taylor, A., Dougill, A. J., Antwi-Agyei, P., Adefisan, E. A., Ahiataku, M. A., ... & Toure, A. (2020). Exploring the need for developing impact-based forecasting in West Africa. *Frontiers in Climate*, *2*, 565500. https://doi.org/10.3389/fclim.2020.565500

Obermeier, H. B., Berry, K. L., Klockow-McClain, K. E., Campbell, A., Carithers, C., Gerard, A., & Trujillo-Falcón, J. E. (2022). The creation of a research television studio to test probabilistic hazard information with broadcast meteorologists in NOAA's Hazardous Weather Testbed. *Weather, Climate, and Society*, *14*(3), 949-963. https://doi.org/10.1175/WCAS-D-21-0171.1

OCHA. (2022). Flash Appeal Malawi. Retrieved from https://reliefweb.int/sites/reliefweb.int/files/resources/ROSEA_20220224_Malawi_TropicalStorm_Ana_Flash_Appeal_Feb-Apr-2022_final.pdf

Patwary, M. M., Bardhan, M., Haque, M. A., Moniruzzaman, S., Gustavsson, J., Khan, M. M. H., ... & Islam, M. A. (2024). Impact of extreme weather events on mental health in South and Southeast Asia: A two decades of systematic review of observational studies. *Environmental research*, 118436. https://doi.org/10.1016/j.envres.2024.118436

Peng, L., Zhang, W., Wang, X., & Liang, S. (2019). Moderating effects of time pressure on the relationship between perceived value and purchase intention in social E-commerce sales promotion: Considering the impact of product involvement. *Information & Management, 56*(2), 317-328. https://doi.org/10.1016/j.im.2018.11.007

Persaud, N., McLeod, P., & Cowey, A. (2007). Post-decision wagering objectively measures awareness. *Nature neuroscience*, *10*(2), 257-261. https://doi.org/10.1038/nn1840

Potter, S., Harrison, S., & Kreft, P. (2021). The benefits and challenges of implementing impact-based severe weather warning systems: Perspectives of weather, flood, and emergency management personnel. *Weather, climate, and society*, *13*(2), 303-314. https://doi.org/10.1175/WCAS-D-20-0110.1

Robbins, J., & Titley, H. (2018). Evaluating high-impact precipitation forecasts from the Met Office Global Hazard Map (GHM) using a global impact database. *Meteorological Applications, 25*(4), 548-560. https://doi.org/10.1002/met.1720

Röösli, T., Appenzeller, C., & Bresch, D. N. (2021). Towards operational impact forecasting of building damage from winter windstorms in Switzerland. *Meteorological Applications, 28*(6), e2035. https://doi.org/10.1002/met.2035

Sai, F., Cumiskey, L., Weerts, A., Bhattacharya, B., & Haque Khan, R. (2018). Towards impact-based flood forecasting and warning in Bangladesh: A case study at the local level in Sirajganj district. *Natural Hazards and Earth System Sciences Discussions*, 1-20. https://doi.org/10.5194/nhess-2018-26

Saintonge, C. D. D., Kirwan, J. R., Evans, S. J., & Crane, G. J. (1988). How can we design trials to detect clinically important changes in disease severity?. *British journal of clinical pharmacology*, *26*(4), 355-362. https://doi.org/10.1111/j.1365-2125.1988.tb03392.x

Silvestro, F., Rossi, L., Campo, L., Parodi, A., Fiori, E., Rudari, R., & Ferraris, L. (2019). Impact-based flash-flood forecasting system: Sensitivity to high resolution numerical weather prediction systems and soil moisture. *Journal of Hydrology*, *572*, 388-402. https://doi.org/10.1016/j.jhydrol.2019.02.055

Singhal, A., Raman, A., & Jha, S. K. (2022). Potential use of extreme rainfall forecast and socio-economic data for impact-based forecasting at the district level in Northern India. *Frontiers in Earth Science*, *10*, 846113. https://doi.org/10.3389/feart.2022.846113

Singmann, H., & Kellen, D. (2019). An introduction to mixed models for experimental psychology. In *New methods in cognitive psychology* (pp. 4-31): Routledge.

Speight, L., Cole, S. J., Moore, R. J., Pierce, C., Wright, B., Golding, B., . . . Ghimire, S. (2018). Developing surface water flood forecasting capabilities in Scotland: An operational pilot for the 2014 Commonwealth Games in Glasgow. *Journal of Flood Risk Management, 11*, S884-S901. https://doi.org/10.1111/jfr3.12281

Spruce, M. D., Arthur, R., Robbins, J., & Williams, H. T. (2021). Social sensing of high-impact rainfall events worldwide: a benchmark comparison against manually curated impact observations. *Natural Hazards and Earth System Sciences, 21*(8), 2407-2425. https://doi.org/10.5194/nhess-21-2407-2021

Stewart, T. R., Heideman, K. F., Moninger, W. R., & Reagan-Cirincione, P. (1992). Effects of improved information on the components of skill in weather forecasting. *Organizational Behavior and Human Decision Processes, 53*(2), 107-134. https://doi.org/10.1016/0749-5978(92)90058-F

Stewart, T. R., Moninger, W. R., Brady, R. H., Merrem, F. H., Stewart, T. R., & Grassia, J. (1989). Analysis of expert judgment in a hail forecasting experiment. *Weather and Forecasting, 4*(1), 24-34. https://doi.org/10.1175/1520-0434(1989)004<0024:AOEJIA>2.0.CO;2

Stewart, T. R., Roebber, P. J., & Bosart, L. F. (1997). The importance of the task in analyzing expert judgment. *Organizational Behavior and Human Decision Processes, 69*(3), 205-219. https://doi.org/10.1006/obhd.1997.2682

Taramelli, A., Valentini, E., & Sterlacchini, S. (2015). A GIS-based approach for hurricane hazard and vulnerability assessment in the Cayman Islands. *Ocean & Coastal Management, 108*, 116-130. https://doi.org/10.1016/j.ocecoaman.2014.07.021

Taylor, A. L., Kause, A., Summers, B., & Harrowsmith, M. (2019). Preparing for Doris: Exploring public responses to impact-based weather warnings in the United Kingdom. *Weather, climate, and society*, *11*(4), 713-729. https://doi.org/10.1175/WCAS-D-18-0132.1

Taylor, A., Summers, B., Domingos, S., Garrett, N., & Yeomans, S. (2024). The effect of likelihood and impact information on public response to severe weather warnings. *Risk analysis*, *44*(5), 1237-1253. https://doi.org/10.1111/risa.14222

Taylor, R. L., & Wilsted, W. D. (1974). Capturing judgmental policies: A field study of performance appraisal. *Academy of Management Journal, 17*(3), 440-449. https://doi.org/10.5465/254648

Trujillo-Falcón, J. E., Reedy, J., Klockow-McClain, K. E., Berry, K. L., Stumpf, G. J., Bates, A. V., & LaDue, J. G. (2022). Creating a communication framework for FACETs: How probabilistic hazard information affected warning operations in NOAA’s Hazardous Weather Testbed. *Weather, Climate, and Society*, *14*(3), 881-892. https://doi.org/10.1175/WCAS-D-21-0136.1

Uccellini, L. W., & Ten Hoeve, J. E. (2019). Evolving the National Weather Service to build a weather-ready nation: Connecting observations, forecasts, and warnings to decision-makers through impact-based decision support services. *Bulletin of the American Meteorological Society*, *100*(10), 1923-1942. https://doi.org/10.1175/BAMS-D-18-0159.1

United Nations. (2015). Transforming our World: The 2030 Agenda for Sustainable Development. Retrieved from https://sdgs.un.org/2030agenda

Wei, L., Li, J., & Yang, X. (2018). Experiments on impact-based forecasting and risk-based warning of typhoon in China. *Tropical Cyclone Research and Review, 7*(1), 31-36. https://doi.org/10.6057/2018TCRR01.04

Weyrich, P., Scolobig, A., Bresch, D. N., & Patt, A. (2018). Effects of impact-based warnings and behavioral recommendations for extreme weather events. *Weather, climate, and society*, *10*(4), 781-796. https://doi.org/10.1175/WCAS-D-18-0038.1

White, N., Harries, P., Harris, A. J., Vickerstaff, V., Lodge, P., McGowan, C., ... & Stone, P. (2018). How do palliative care doctors recognise imminently dying patients? A judgement analysis. *BMJ open*, *8*(11), e024996. https://doi.org/10.1136/bmjopen-2018-024996

Whiteley, L., & Sahani, M. (2008). Implicit knowledge of visual uncertainty guides decisions with asymmetric outcomes. *Journal of vision, 8*(3):2, 1-15. https://doi.org/10.1167/8.3.2

Wilkinson, S., Dunn, S., Adams, R., Kirchner-Bossi, N., Fowler, H. J., González Otálora, S., . . . Chan, S. C. (2022). Consequence forecasting: A rational framework for predicting the consequences of approaching storms. *Climate Risk Management, 35*, 100412. https://doi.org/10.1016/j.crm.2022.100412

Wilsted, W. D., Hendrick, T. E., & Stewart, T. R. (1975). Judgement policy capturing for bank loan decisions: An approach to developing objective functions for goal programming models. *Journal of Management Studies*, *12*(1-2), 210-215. https://doi.org/10.1111/j.1467-6486.1975.tb00895.x

WMO. (2015). WMO Guidelines on Multi-hazard Impact-based Forecast and Warning Services. Retrieved from https://library.wmo.int/doc_num.php?explnum_id=7901

WMO. (2021a). Water-related hazards dominate disasters in the past 50 years. Retrieved from https://public.wmo.int/en/media/press-release/water-related-hazards-dominate-disasters-past-50-years

WMO. (2021b). WMO Guidelines on Multi-hazard Impact-based Forecast and Warning Services. Part II: Putting Multi-hazard IBFWS into Practice. Retrieved from https://library.wmo.int/doc_num.php?explnum_id=10965

World Economic Forum. (2023). The Global Risks Report 2023 18th Edition. Retrieved from https://www3.weforum.org/docs/WEF_Global_Risks_Report_2023.pdf

Wyatt, F., Robbins, J., & Beckett, R. (2023). Investigating bias in impact observation sources and implications for impact-based forecast valuation*. International Journal of Disaster Risk Reduction, 90*, 102339. https://doi.org/10.1016/j.ijdrr.2023.103639

Yu, J., Liu, J., Baek, J.-W., Fong, C., & Fu, M. (2022). Impact-based forecasting for improving the capacity of typhoon-related disaster risk reduction in typhoon committee region.

Tropical Cyclone Research and Review, 11(3), 163-173. https://doi.org/10.1016/j.tcrr.2022.09.003

Zsambok, C. E., & Klein, G. (2014). *Naturalistic decision making*: Psychology Press.

Appendix

Appendix 1

Table A1

Impact list summarized from qualitative surveys of partner countries

Death - The number of reported deaths	Class / work suspension/daily activities
Injury - The number of people injured	Increase road accidents
Miss - The number of people missing	Stranded passengers
Displace - The number of people displaced (made homeless)	The time of the event (during the after-office hour / peak time)
Evacuate -The number of people evacuated	The number of municipalities affected
Affect - The number of people affected	Infrastructure damaged (road, bridge, house)
Building Damaged - The number of buildings damaged	Damage to property - utilities (power plant/electricity pole)
Hospital - The number of hospitals or health centres affected	Damage to the dam wall / River embankments
School - The number of educational facilities affected	Road damaged
Public building - The number of government, public or religious buildings affected	Damage to water pipes / sewerage system
Transport Infrastructure -The number of transport infrastructure assets affected	Damage to property - residential/house
Road Sections - The number of road sections affected	Damage to public facilities
The length of recovery after the event (e.g. if utilities were damaged, how long before they were back)	The areal extent - how widespread is the effect / damage in a certain locality/municipality
Bridges - The number of bridges damaged or affected	Road Distance (Km) - The distance of road affected
Agriculture and aquaculture -The total area of agriculture and aquaculture affected	The duration of the event
Livestock - The total number of agricultural animals and poultry killed	Spread of other health issues (cough, fever, dengue) / infectious disease
Waterborne diseases due to prolonged flooding	The location of the event
Water contamination	Cause trauma and escalate the distress
Plants were damaged	Flood / Flash flood / water pooling
Soil erosion	Increase in river/creek level heights
Food supply	Landslide
Communication failure/Isolated community	Lightning strikes
Water supply	Rock fall
Disrupt transport	Wet roads and reduced visibility
Traffic congestion	Debris / mud flows
Flight/train cancellation/delay	Airport closure
Difficult to drive vehicle on the street	Disruption at the port / waterways
Energy supply	

Appendix 2

Ensuring the ecological validity of the heavy rainfall events used – database analysis and simulation

Estimating the multivariate distribution of rainfall events from the database and generating hypothetical rainfall events from it required several steps. Firstly, the database initially contained entries on an individual record level and not on an event level; that is, the same event could have multiple entries (e.g., from different data bases or different days). Therefore, we first needed to reduce the database so that only one entry per weather event remained. We used two different reduction approaches, a strict approach and a less strict approach. In the strict approach we matched events based on the city level and consecutive days, in the less strict approach we matched events based on country and a time window of 7 days. In the strict approach we ended up with a total of around 7,500 events and in the less strict approach with a total of around 4,000 events. If for one event different entries had values recorded for the same impact, we used the maximum value across entries as value for this event.

In the second step, we explored the distribution of the impacts for the two event data sets. This data exploration provided several important insights. For both approaches, the vast majority of events did not have values for all impacts. For some impacts the number of missing values was comparatively low, (e.g., for deaths, the proportion of missing values was 43% for the strict approach and 7% for the less strict approach) and for other impacts we had hardly any values (e.g., for roads and bridges the proportion of missing values was 97% and 98%). When looking at the univariate distributions of each impact, all were extremely right skewed with most having some extreme outliers. To handle the extreme outliers in each distribution we winsorised each univariate distribution at the 99th percentile. To address the extreme right skew we decided to log-transform the data. Because the deaths variable also included many events with 0 deaths, we used a log1P transformation which is defines as $x_{\text{transformed}} = \ln{(1 + x)}$. The histograms in the main diagonal of Figure 3 show the univariate distribution of the impacts from event data based on the less strict approach after the log1P transformation. As can be seen, for some of the impacts (i.e., Affected, Houses, and Agriculture), the univariate distributions of impacts are approximately normal after transformation. For the remaining impacts there still is a noticeable right skew after transformation, but it is much less dramatic compared to pre-transformation.

In the third step, we estimated the multivariate normal distribution of the log1P transformed impact variables in a Bayesian statistical framework (using Stan) (Carpenter et al., 2017). Because of the large amount of missing data, we estimated the multivariate normal distribution in such a way that each event informed the part of the multivariate normal distribution for which it did have information. For example, an event that only had data for one impact, only informed the mean for that impact type. An event with data for two impacts, informed the corresponding two means as well as the covariance between both impacts. Because we had two different event data sets (one based on the strict and one based on the less strict event reduction approach), we estimated the multivariate normal distribution jointly on both data sets. This joint analysis was chosen to alleviate concerns regarding the specific choices taken in the event reduction step. We felt the downside of the joint analysis – it can be seen as an instance of pseudoreplication (Hurlbert, 1984) – were less of a concern given the relatively large sample sizes. The analysis resulted in one posterior distribution for the multivariate normal distribution over all six impacts on the log1P transformed scale.

In the fourth step, we drew 1,000 samples from the multivariate posterior distribution (i.e., each sample represents one set of parameters for the full multivariate normal distribution). From each of these samples, we generated two hypothetical weather events using a random number generator for the multivariate normal distribution. This resulted in a total of 2,000 hypothetical weather events. Because not all hypothetical weather events were generated from the same set of parameters of the multivariate normal distribution, but from different samples of the full posterior distribution, the simulation ensured that the uncertainty we had in the parameters of the multivariate normal distribution (i.e., the means and covariances of the variables) were properly represented in the hypothetical data. Because the hypothetical weather events were on the log1P transformed scale, we back-transformed them onto the actual event scale using $(e^x - 1)$ and then rounded them to whole numbers. Furthermore, all negative impacts were set to 0. We then checked the resulting events for certain patterns that could occur in the hypothetical events, but were unlikely to occur in real data. For example, in 67 of the hypothetical 2,000 events the number of affected people was less than two times the number of affected houses, which we deemed unlikely. Likewise, there were 21 events with more affected public buildings than affected houses. We removed these events from further consideration. The upper triangle of Figure 3 shows the distribution of the remaining hypothetical weather events, after transforming them back to the log1P scale to deal with the strong right-skew. The univariate as well as multivariate pattern of the hypothetical events strongly matches the characteristics of the real weather events.

Appendix 3

Table A2

Impact values used in Impact Threshold Task

Impacts	People dead	People affected (e.g., injured, displaced, evacuated)	Houses damaged or destroyed	Public buildings affected (e.g., schools, hospitals, government or religious buildings)	Agriculture / aquaculture affected (hectares)	Road sections and / or bridges closed
Number1	1	38	3	1	5	1
Number2	2	(39, 244)	(4, 20)	2	(6, 32)	2
Number3	5	(244, 532)	(20, 43)	5	(32, 65)	5
Number4	(6, 8)	(532, 932)	(43, 74)	10	(65, 108)	(6, 10)
Number5	(8, 15)	(932, 1505)	(74, 117)	(11, 18)	(108, 167)	(10, 16)
Number6	(15, 25)	(1505, 2355)	(117, 178)	(18, 31)	(167, 250)	(16, 24)
Number7	(25, 61)	(2355, 3685)	(178, 272)	(31, 48)	(250, 374)	(24, 52)
Number8	N/A	(3685, 5948)	(272, 427)	(48, 108)	(374, 576)	N/A
Number9	N/A	(5948, 10417)	(427, 723)	N/A	(576, 953)	N/A
Number10	N/A	(10417, 22661)	(723, 1501)	N/A	(953, 1917)	N/A
Number11	N/A	(22661, 143484)	(1501, 8488)	N/A	(1917, 10063)	N/A

Note. The yellow cells represent the fixed numbers. The green cells represent drawing a random number from the range. (A, B): A is inclusive, B is exclusive.

Appendix 4

Table A3. Trend estimates and their standard errors (in parentheses) of six impact predictors for four SEA countries in the Impact Model.

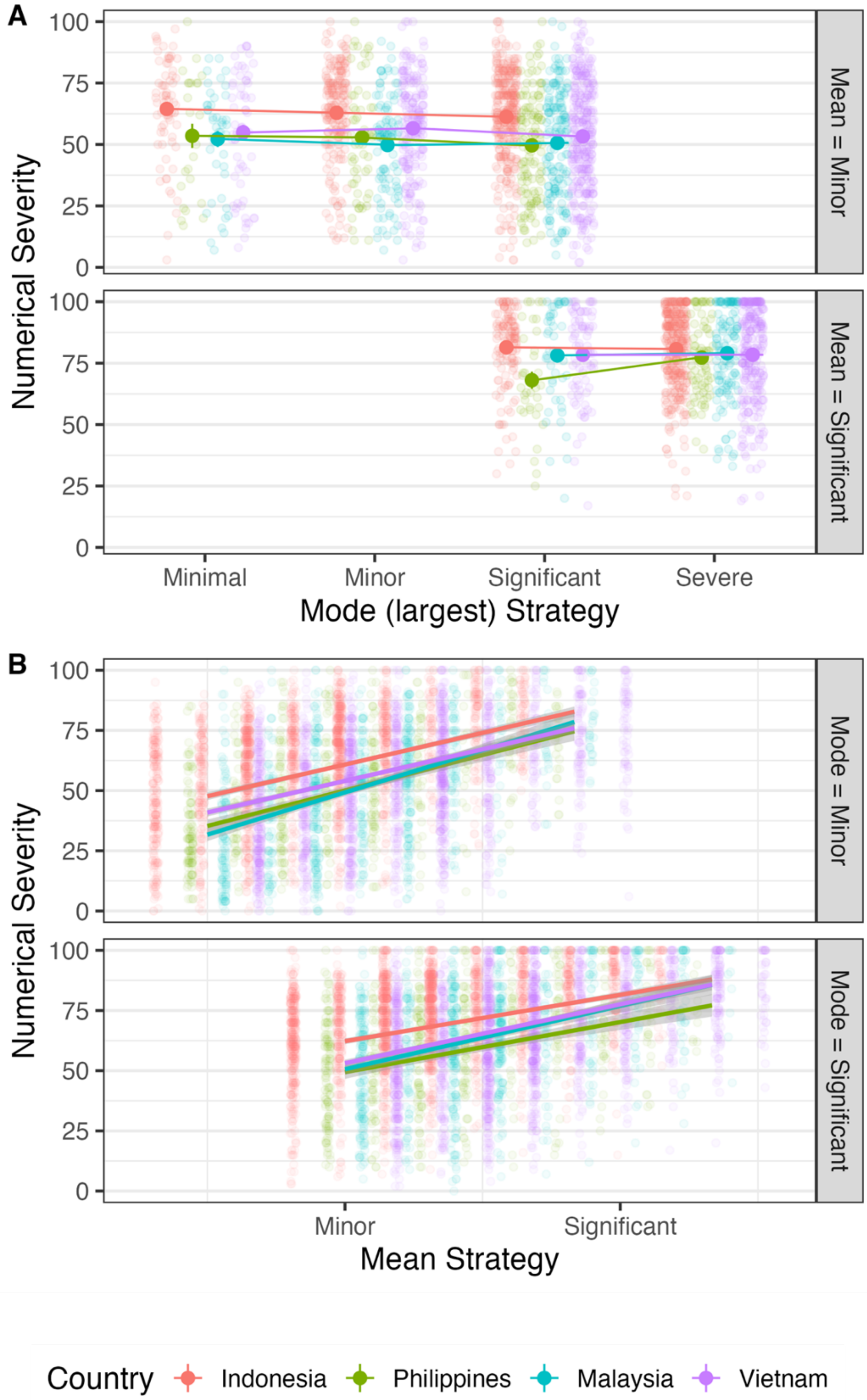
Predictor	Indonesia	Philippines	Malaysia	Vietnam
Deaths	3.44*** (0.41)	4.09*** (0.73)	4.98*** (0.62)	6.17*** (0.43)
Affected	6.09*** (0.56)	6.92*** (1.00)	9.23*** (0.85)	6.76*** (0.59)
Houses	2.15*** (0.36)	2.74*** (0.64)	2.24*** (0.55)	1.56*** (0.38)
Public buildings	1.17*** (0.23)	1.03* (0.42)	0.72* (0.35)	0.84*** (0.24)
Agriculture/ Aquaculture	2.03*** (0.22)	3.53*** (0.40)	2.10*** (0.34)	1.64*** (0.23)
Roads/ Bridges	0.96*** (0.25)	2.05*** (0.45)	1.26*** (0.37)	1.16*** (0.26)

Note. *** *p* < .001; ** *p* < .01; * *p* < .05.

Appendix 5

Figure A5.1.

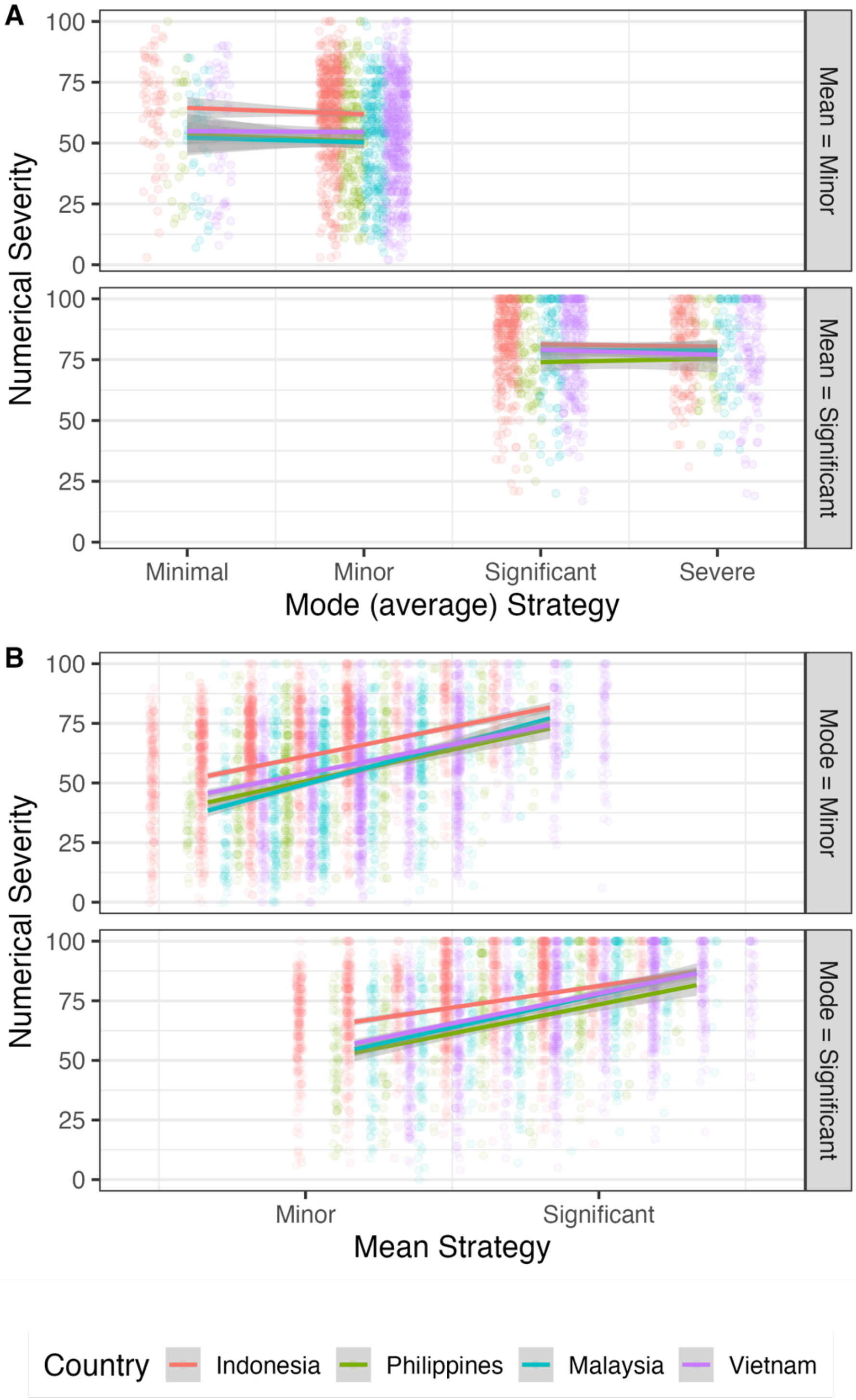
Comparison between Mean Strategy and Mode (largest) Strategy



Note. In each panel, different colours represent different countries, with each datapoint reflecting one observation. Datapoints are plotted semi-transparently to avoid overplotting, such that darker points do indicate more data. In Panel A, individual data points are additionally jittered randomly on the x-axis. Furthermore, the large points represent means (with error bars representing the standard error of the mean). In Panel B, the solid lines show the regression lines.

Figure A5.2.

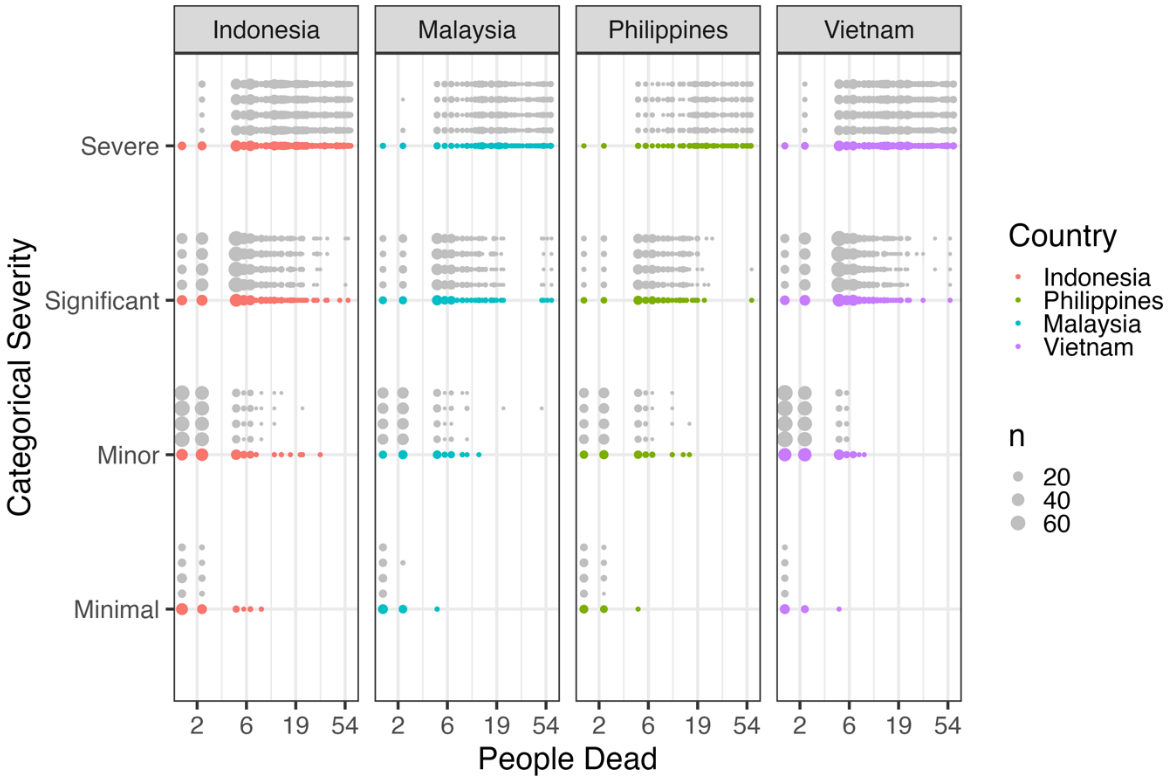
Comparison between Mean Strategy and Mode (average) Strategy



Note. See Figure A5.1.

Appendix 6

The plot comparing the observed and predicted categorical severity ratings for the impact of People Dead in Individual Threshold Task. Plots for each of six impact types are available in OSF. Note that the colourful dots are observed data and the grey dots are predicted data.


Indonesia
Malaysia
Philippines
Vietnam
Categorical Severity
Severe
Significant
Minor
Minimal
People Dead
Country
Indonesia
Philippines
Malaysia
Vietnam
n