Standardizing Catastrophe Model Validation

Contents

Introduction

Catastrophe models were first introduced in the late 1980s in order to quantify the risk of financial loss to portfolios of insured properties. Their main output is a summary of the average annual loss to the portfolio (also known as the 'expected loss') and an array of larger loss sizes that are expected to be exceeded at a given annual probability. The '1 in 100' year loss for example refers to the size of loss likely to be exceeded in any given year with a 1% probability. All models are imperfect representations of reality, however, and directly reflect the expertise, experience and data used to build them. As a result, they can vary dramatically in their loss outputs. The author has seen models that vary by as much as twenty times in the size of loss that can be exceeded at a return period of 1 in 200 years. It is obviously important to be able to differentiate between such models, specifying which ones appear reasonable and which ones do not. This is achieved by a process known as 'model validation'. Unfortunately, this term is loosely defined and can mean widely differing things to different people, ranging from republishing of vendor documentation, through version change analysis, to 'deep diving' into the inner workings of a model to see why it has changed. There is no formal definition of model validation, and no set of standards governing which activities should be included and which should be excluded. As a result, model validations are often variable in scope, quality, depth and insight provided. This paper argues for the introduction of a standard definition and approach to model validation, to improve the quality and streamline the efficiency of the process.

The primary consumers of model validation reports tend to be financial services regulators, internal governance bodies and pricing actuaries. Solvency II places the onus firmly on the model licensee to understand the risk posed and the models used to represent that risk. Internal catastrophe risk management committees also typically seek validation-based insight into new major peril models that may be adopted, or where a major peril model has changed significantly between versions. Pricing actuaries are usually keenly interested in developing their 'own views of risk' to superimpose on the catastrophe model output, to change the frequency/ severity of selected events and/ or to correct for missing subperils. Model validation can lay an appropriate foundation for all of these activities.

Defining model validation

A useful starting definition for model validation is available from the Lloyd's Market Association: '***Validation is … the process by which you determine whether the external catastrophe model provides a valid representation of the catastrophe risk for your portfolio'*** (LMA, 2012). This explicitly refers to the importance of assessing the suitability of the model for use with the user's own portfolio. This is not a complete picture, however, because it is possible for a model:

1. To be of use with an industry portfolio (against which it is calibrated, a large chunk of which comprises single family dwellings) but a poor fit to a user's portfolio (such as a portfolio of large power utilities).
2. Conversely, to be of use with a user's portfolio but a poor fit at industry level.
3. To represent the peril it is representing well (in terms of hazard) but be a poor fit to portfolios analysed (in terms of loss).

This means that cat models can be better suited for some purposes and less well suited for others. We should therefore start by asking two fundamental questions:

1. **Is the model a good fit against data likely to have been used to build/ calibrate it?** This is concerned with assessing the 'goodness of fit' of the model and its components to historic data known or likely to have been used to build (such as a prominent event catalogue) and calibrate it (such as the vendor's industry exposure database and a string of major well known industry loss events). It says little to nothing about the suitability of the model for use with a different portfolio, however.

2. **Does it have predictive skill with data not used to build/ calibrate it?** This focuses on the skill of the model in regard to the user's own portfolios, which were not used to calibrate the model during development. Skill can be assessed using commonly available yardsticks, including the portfolio's historic loss record and modelled scenario losses. Such techniques are routinely employed in actuarial validation of Internal Capital Models and should also be applied to catastrophe model validation. Using these techniques, the author found it straightforward to disqualify several vendor models on the basis that recent historic losses plot in the distant tail of their loss distribution on specific industry-similar portfolios in Asia-Pacific, in areas known to be amongst the world's most exposed to earthquakes and typhoons.

Without such a two-part definition, it is common for model validators to spend the majority of their time answering question one. The risk here is that by addressing question one, we may also erroneously conclude that question two is also true, i.e., that a model must also be fit for use with our own portfolios. This is not automatically the case, and hence needs to be established. Model validators should therefore seek to answer both questions, and as far as possible should look to do so in roughly equal measure in terms of time and effort spent.

Implementation

Seeking to answer core questions 1 and 2 above does not mean that we should spend equal time on every catastrophe model, however. Some models generate much higher losses to our portfolios than others, and hence are considered to be more 'material' than others. We should spend more validation time and effort on models that are more material, and less (or even none) on those that are not. Table 1 provides an illustrative example of how materiality could in principle be mapped to the model validation exercise to be undertaken. A high materiality peril for example might cause a 1 in 100 year loss exceeding $250 million and/ or contribute more than 20% of gross premium income across all lines of business. This would map in this example to a 'Full' validation.

Table 1. Example of mapping between materiality and validation category via two alternate materiality metrics (OEP 1:100 Value at Risk (VaR) and % of GWP).

OEP 1:100 VaR (USD Millions)	**% of Gross Written Premium (GWP)**	**Materiality Category**	**Model Validation Category**
>250	>20%	High	Full
50-250	>2<=20%	Low	Basic
<50	<=2%	Immaterial	Light touch, or None

Having decided on the category of model validation to be undertaken based on the materiality of the peril in question, the next step is to tailor the validation activities to fit. Table 2 lists activities that may be appropriate within each of the validation categories shown in Table 1, grouped into 'validation modules'.

Table 2. Proposed validation categories and modules, and examples of their content.

Validation Category	**Validation Module**	**Example of Content**
FULL	Peril outline	- Description of the peril, without regard for how it is implemented in the model.
	Loss validation	- Industry back-testing, scenario testing, historic loss comparison (modelled vs. actual). - Own portfolio back-testing, scenario testing, historic loss comparison (modelled vs. actual). - Model parameter sensitivity testing (choice of event catalogue, demand surge, sub-peril impact, etc.). - Summary of findings.
	Hazard validation	- Overview of peril hazard implementation in the model. - Event rate benchmarking by sub-peril, by region. - Event footprint intensity benchmarking by sub-peril, for selected events/ event sources. - Summary of findings.
	Vulnerability validation	- Portfolio level sensitivity to changes in primary & secondary exposure attributes relevant to the portfolios being analysed. - Location level reconstruction of damage curves - Testing for regionalization in vulnerability modelling. - Summary of findings.
	Geocoding	- Portfolio impact of disaggregation of aggregate exposures. - Portfolio impact of geocoding at different spatial resolutions. - Summary of findings.
BASIC	Peril outline	- As above, drawn from vendor documentation.
	Loss validation	- Reuse existing vendor industry loss validation work and add back-testing of industry loss experience from own testing. Also consider whether to add back-testing of own portfolio loss experience.
LIGHT TOUCH	Peril outline	- As above, drawn from vendor documentation.
	Loss validation	- Simply reuse existing vendor industry loss validation work.
APPLIES TO ALL	Background	- Objectives to be fulfilled (core questions 1 & 2). - Summary of relative peril materiality. - How and why the model for this specific peril region is used in the business. - Commonly held market views on the model and its results.

Table 2. Proposed validation categories and modules, and examples of their content.		
Validation Category	**Validation Module**	**Example of Content**
		- Any specific reasons for investigating this peril model.
	Model & portfolio metadata	- Model geographic scope, subperil completeness & capabilities. - Model analysis settings used. - Portfolios used (summary of) and any indexation applied.
	Summary	- Summary of model pros, cons, errors & omissions, and key dependent assumptions made, grouped by validation module, and ranked in priority order of size of impact on model loss outputs.
	Conclusions	- High level recommendations on adjustments needed to correct for errors & omissions. - Recommendations on suitability of the model in context of core validation questions 1 & 2, in respect of the issues identified above (if any). Recommendations range from (a) accept 'as is', (b) accept with adjustment, or (c) reject.
	Independent review	- Commission an independent review of the validation report for feedback and assurance prior to the sign off stage.
	Sign off	- Record of the sign off process for (a) the validation report and (b) operational use of the model.

Once the mapping between materiality, validation category and module has been decided, the content of each module then needs to be fleshed out. The content should match the validation category, for example a 'full validation would contain all the loss validation module content outlined in Table 2, whilst a 'basic' validation may contain only a small number of selected items. Once the content of each validation module has been decided, it is then a matter of creating the project plan and engaging appropriate resources to fulfil it.

In practice, it is inevitable that there will often be blurring of the boundaries between validation categories because of 'scope creep', in which an exercise originally intended to be 'light touch' may end up as a 'basic' validation, etc. This is not seen as a problem from a validation standpoint, because an almost inevitable result is that such blurring will result in more detail being collected about the model and its use, rather than less. Blurring and scope creep is something for model validation project managers to be aware of and monitor, however.

A last point here is that knowing when to pause or cease validation activities (potentially mid-way through the project) is just as important in knowing what to do. Note that loss validation is the only major technical module that is common to all the validation categories in Table 2, and is intended to be performed prior to the hazard, vulnerability and geocoding modules. If a model is judged by the validation team to fail loss validation testing (i.e., it severely understates or overstates the arrival rate of losses of certain sizes, or the results fall outside a broad range of acceptability by the market[1]) a decision will be needed on whether to continue with the validation in knowledge that an adjustment is needed to be subsequently developed (as is likely where the under- or over-statement is not too dramatic or material), or to pause or even cease the validation and seek a suitable explanation from the vendor.

Preferred Exclusions

Wider business stakeholders often consider the following items to be integral to model validation studies; however these topics appear incongruous for the reasons outlined below:

[1] For example, if a loss of the size of Hurricane Andrew (normalized to present day values) plotted at only five years on the current industry loss curve. A more realistic range for a loss of this size is roughly 15-30 years.

1. **Developing an 'own view of risk':** This is a user-defined view of the arrival rate of losses of certain sizes superimposed on the default reference view of risk from the model vendor. Such views need to be informed by and based on the model validation so that they benefit from the knowledge gained and do not contradict the validation conclusions. This means that a view of risk should be developed after the model validation is complete, rather than prior to or during. View of risk development also requires input from a much broader team of contributors (including pricing actuaries, exposure management professionals and underwriters) than the model validation team; and will also be updated more regularly (validation exercises are typically undertaken at multi-year intervals, whilst the view of risk is likely to be iteratively adjusted as needed). It therefore makes sense to develop an own view of risk once the validation is complete.
2. **Verifying model functionality and financial calculations:** It is important to be confident that model functionality is working as intended, and that financial loss metrics produced are generated in a consistent manner that is accepted by the market. Discrepancies can arise in particular in the application of primary policy conditions (e.g., whether a deductible should be applied before or after the policy limit) and with reinsurance conditions (e.g., whether layer recoveries under specified hours clauses should be optimised by layer or by programme). Confirming that functionality and financial calculations work as intended in the model is a task that should be performed once only, ideally by the model vendor or modelling framework provider. However, in practice it is clear that vendors do not necessarily agree, because models are released that lack intended features or contain features that simply do not function as expected. This is especially likely with new model platform releases that have not yet been subjected to extensive live use by the vendor's client base. Early detection and correction of such errors is important; however this could easily consume the entire time allocated to model validation activities and detract from answering the core validation questions outlined above. Verification of model functionality is therefore identified here for exclusion from the model validation process. Instead, functional and financial module verification should be addressed separately, once only for each model version release, preferably by an industry body set up or adapted to perform this important task.
3. **Impact of future climate change:** The extent to which historic and recent climatic change influences extreme event occurrence rates and severities clearly lies within the scope of model validation. By contrast, the potential impact of speculative future climate change on rates and severities shorter than several decades into the future is not obviously within scope of a validation exercise designed to assess the suitability of use of a model under today's climate. Future climate change impacts are themselves inherently uncertain, let alone the impact of such changes on the rates and severities of extreme events embedded within possible future climate states. Our ability to robustly model such phenomena is also limited by the capabilities of regional and global climate models, which display skill only when used over timescales longer than several decades and over sub-continental scales and greater (Fiedler et al., 2021). Therefore, important as it is, assessing the impact of future climate change on the modelled peril being investigated should constitute a separate follow-on study.

Other Potential Improvements

The following items also need to be considered when designing and implementing the validation study.

1. **Establish design standards for the structure and content of model validations:** This report provides pointers to ways in which the structure, content and quality of model validation exercises can be improved. It is recommended that a standard set of procedures and templates be agreed and published freely online, to promote quality, consistency and depth of modelling insights gained.

2. **Develop minimum standards for validation documentation:** It is essential to 'know your customer' in advance of undertaking a model validation exercise. Non-technical business stakeholders for example commonly prefer a short presentation and/ or short factsheet to be given to them that summarizes the key findings of the study. Technical stakeholders and business managers however will likely demand much more depth and insight. For quality assurance and record keeping purposes, there should also be a means of bringing all the detailed results of the study together into a single 'document of record'. All of this means that there will be multiple levels and types of documentation needing to be produced, some of which can be updated regularly as the project proceeds, and some of which will be generated after the project has concluded. It is proposed that minimum standards for documentation be defined that are suitable for consumption by a range of different stakeholders.

3. **Pass/fail approach may need to be replaced:** Actuaries in particular often prefer binary pass/ fail test results. To implement this, we need to know in advance what the result should be in order to define the pass criterion. Unfortunately, in catastrophe modelling it is often impossible to define a definitive pass criterion, because no one knows the answer in many cases. The alternative is to build a qualitative case for the model either appearing correct or incorrect, at all times acknowledging that the answer is not actually known. For example, model 'A' may assume an arrival rate of 20 years for loss-causing tsunamis at the U.S. Pacific coast. We know that the Port of Los Angeles has existed in its present location for at least 100 years, and that it was damaged only once by the tsunami in 1964 from the Great Alaskan earthquake. The known frequency of arrival of damaging tsunami events is therefore 1 in 120 years (0.83% p.a.), which is much less than the modelled frequency of 5% p.a. This would instantly show that the modelled rate appears overstated. Note that neither rate was known at the start of the process, hence it would only have been possible to define a pass/fail test in retrospect.

4. **Model validation datasets should be standardised and curated:** It is desirable to define a standard minimum palette of datasets that should at least be considered when performing a model validation study, to promote consistency and completeness between validation practitioners. The master list of datasets would need to be constructed in light of specific objectives. For example, to support loss validation it would be extremely useful to be able to freely access a peer-reviewed open dataset of historic industry insured losses per peril by geography (both in 'as was' dollar values and on-levelled to present day values). To support hazard validation, an open dataset of historic event arrivals would be specified for each peril (updated regularly to capture ongoing re-analysis work and recent historical events); whilst for vulnerability, the source of the building structural design code and a dataset of expected lateral loading forces at a given return period (e.g., from wind or ground shaking) would be outlined.

5. **Increase model vendor involvement:** Experience has shown that the majority of model vendors are disincentivised to find public fault with their own models. This does not mean that they should have no further role in model validation, however. Rather, model vendors are uniquely placed to be able to test the sensitivity of their models to changes in some of the core parameter values on which their default 'reference' view depends; this is a task that simply cannot be achieved without their involvement. Model vendors could also be tasked with creating a generic model capabilities proforma; for industry-level loss validation; and for producing key datasets that are routinely used in validation (such as a gridded AAL or hazard intensity map), instead of leaving model licensees to generate these datasets themselves. Specific ways in which vendors could assist include the following:

    a. **Model capabilities proforma:** Listing the scope and capabilities of the model is a basic introductory task in every model validation study. At present, model validators typically need to scour vendor documentation for the required data and/ or ask the vendor's technical support team for specific assistance. It would be much more efficient if the vendor itself were to complete a short standard proforma listing the minimum specification of the model (e.g., release/ update history, geographic scope, subperils included/ excluded, catalogues available, vulnerability views available, whether the model assumes flood defences or not, etc.) that it would then make available to clients and regulators, and update as needed with new model version releases.

b. **Standardised industry-level loss validation:** Model vendors often selectively undertake back-testing, scenario testing and historic vs. modelled loss comparisons on an industry portfolio, published in their technical model documentation. It is recommended that this be put on a more formal basis, in which at every new model version release the vendor performs a standard array of industry-level loss validation tests using their prior and current industry exposure portfolios, normalised to present day values. This would eliminate much of the need for model licensees to run these tests themselves; and would show that the model appears reasonable for one of the portfolios that was almost certainly used to calibrate it during development, thereby helping to answer question one at the start of this paper.

c. Production of routine validation datasets: Model validators often generate very similar datasets, e.g., a gridded AAL or hazard intensity maps. It would be much more efficient if model developers produced such resources themselves. In addition, depending on the scale and resolution, some of these datasets can be far too computationally intensive for most validators to produce, for example a 100m wildfire or flood risk grid for the entire contiguous United States. Ideally such products would be offered free-of-charge for model validation purposes; but could be charged-for if used for underwriting purposes, as an incentive for the model vendor.

d. **Functional verification:** The modelling community currently relies on vendor-led internal testing of model functionality and robustness of the associated financial calculations when receiving a new modelling platform, after which it is then ‘beta tested’ by the user community on the vendor’s behalf. This is an obvious systemic industry-wide risk, in that problems and issues may not surface for some time, and when they do, the model may already have been signed-off for operational use. A first step to improving this situation would be for the model vendor to publish (for clients) documentary evidence of the internal testing performed, to allow clients to assess the extent to which the model has been tested prior to release. This is standard behaviour in major bespoke software development projects; but is rarely mentioned in the catastrophe modelling community.

e. **Assessing sensitivity of the model to changes in key assumptions:** The vast majority of cat models generate a single set of financial loss result metrics, such as a loss/ exceedance probability curve and expected loss. These metrics represent the default ’reference’ view from the vendor. Some vendors also provide the option in some models to quantify the impact of ‘high’ and ‘low’ vulnerability states; to use different stochastic event catalogues (e.g., long term rates, medium term rates, warm sea surface temperature, climate conditioned catalogues, etc.); and to add hazard increments at location level to make certain site locations more or less vulnerable. However, few (if any) allow the user to vary event hazard intensity modelling; none allow for simulation of a contingent temporal sequence of dependent subperils (such as aftershocks, fire following outcomes or tsunamis); none provide a detailed presentation of vulnerability relativities; and only a few allow the user to vary the assumed correlation method and weight used when aggregating loss severity distributions from location to portfolio level. Very few vendors provide any standard insight into the impact of adjusting the values of these key variables on reference portfolios, nor proof that their chosen parameterisation is the most appropriate, rather than simply being preferred or the outcome of the modelling. Model vendors should be encouraged to explore and publicise the impact of appropriately adjusting the core parameter values that they have selected during model development. This work would then be made available to clients and their regulators. In the first instance this would be used to support the validation exercise. It could also be used by the vendor as a tranche of alternate views of risk that could be packaged with the default reference view, perhaps for an additional fee as an incentive.

f. **Impact of stochastic catalogue optimisation:** Certain vendor models contain only a few thousand stochastic events in their catalogue, and as such cannot be considered to be fully representative of the hazard posed by the peril to the territory in question. This is justified by the model vendors in terms of optimisation of the catalogue, in which events have been intentionally pruned where this does not significantly alter the shape or valuation of the model loss metrics, in order to reduce model runtime. When attempting to validate the optimised model we therefore often find discrepancies in terms of rates of stochastic events of different intensity/ magnitude intervals between the model and independent benchmark datasets. These differences are to be expected and do not necessarily mean that the model needs adjustment. To reach a conclusion that the model needs to be adjusted would require access to the original un-optimised 'full' catalogue of events. This however is the preserve of the model vendor, which means that model licensees are unlikely to justifiably be able to reach a conclusion about event rates when looking only at the optimised event catalogue. With this in mind, model vendors should be encouraged to document the optimisation process undertaken, including categorisation statistics for stochastic events by intensity both before and after optimisation has been undertaken. Access to such information will allow model validators to assess whether event rates need adjustment.

Proposed Next Steps

The following steps are recommended to standardise, streamline and improve the structure, quality and efficiency of the validation process:

1. Establish a market forum for seeking wider input into this ongoing discussion from heads of modelling research, involving several workshops that highlight internal best practices employed by individual market participants, prior to 'bringing it all together' in a structured format that can be presented to wider stakeholders. An obvious facilitator for this is the Oasis LMF.
2. Once initial consensus has been reached, seek 360° feedback from interested relevant business stakeholders, to include underwriting management, CROs, modellers and pricing / capital model actuaries engaged in primary insurance and reinsurance.
3. Based on this consensus, define minimum standards for the catastrophe model validation process and documentary deliverables that are sensitive to the materiality of the peril region and assumed target audiences.
4. Reach out to supportive regulators once wider agreement is in place, to brief them on the standards agreed and to solicit their further cooperation, e.g., the issuing of guidance to their markets.
5. Significantly increase the contribution from model vendors in regard to transparency around model parameter sensitivity, catalogue optimisation, standardised industry loss validation, production of key datasets and model descriptive proforma, and in verifying the functionality of

their own models. Some of these items could have commercial incentives, for example production of alternate views of risk which could be licensable.

6. Verification of model / modelling framework functionality and production of financial loss metrics is an important area that needs to be addressed to ensure all works as intended. Vendors need to be encouraged to provide documentary evidence of the verification testing performed prior to model release , to their client base and to regulators. It is also proposed here that a bespoke organisation might also be established to undertake independent functional and financial loss verification, with broad industry support. This may be most successful in the first instance with the universe of models based on the open-source Oasis loss modelling framework.

References

Fiedler et al. (2021) Business risk and the emergence of climate analytics. Nature Climate Change, April 2021.

Disclaimer

The information contained in this document is for informational purposes only. While we strive to ensure the accuracy and completeness of the content, we make no guarantees, warranties, or representations about the reliability, suitability, or availability of the information provided. Any action you take based on the information in this document is strictly at your own risk, and we will not be held liable for any losses or damages arising from such actions. The views and opinions expressed are those of the authors and do not necessarily reflect the official policy or position of any organization. Always seek professional advice before making significant decisions.

Please see our Anti-competitive trust policy here: https://oasislmf.org/anti-competitive-trust-policy

Acknowledgements

We wish to thank the following parties and their supporting organisations for their substantive assistance in contributing material and views on which this paper is based: Dickie Whitaker (Oasis LMF), Matt Donovan (Oasis LMF), Tom Philp (Maximum Information), Crescenzo Petrone (Gallagher Re) and Nigel Winspear (Sompo International).