Evaluating Model Analysis of Climate Change Mitigation

Charlie Wilson

*November 2015, IAMC Conference*

(Elmar Kriegler, Jana Schwanitz
Celine Guivarch, Volker Krey, Keywan Riahi, Detlef van Vuuren)
How are IAMs evaluated? To what end? Why is IAM evaluation less visible than climate model evaluation?
Evaluation is about whether models generate the “right behaviour for the right reasons”

**structural validity**

*model is an accurate representation of the system response being modelled*

**behavioural validity**

*model predictions are consistent with observational data*

---

**Behavioural validity** cannot be demonstrated for simulation models of **dynamic**, complex systems.

**Structural validity**

- *model is an accurate representation of the system response being modelled*

**Behavioural validity**

- *model predictions are consistent with observational data*

- **Non-modelled conditions**
  - over-tuning, non-uniqueness
  - limited to historical conditions

---

Structural validity can not be demonstrated for simulation models of dynamic, complex systems.

- **Structural validity**: Model is an accurate representation of the system response being modelled.
- **Behavioural validity**: Model predictions are consistent with observational data.

- Irreducible uncertainties (data, parametric, structural)
- Necessary simplifications

Oreskes 1998 | Schindler & Hilborn 2015
IAM evaluation is an **open-ended process** of testing, learning & improving a model and its performance

**Evaluation criteria for IAMs**

**appropriateness** is model purpose and design consistent with the research question?

**interpretability** are model results clearly interpretable in light of model structure and parameterisation?

**verifiability** are model results repeatable or is model structure accessible to 3rd parties?

**credibility** is model seen as good enough for its intended purpose by both users and modellers?

**usefulness** do model insights help understand uncertainties, trade-offs, alternatives?
Different **evaluation methods** are used with IAMs, particularly to test structural validity.

**structural validity**

- model checks
- transparent documentation
- expert review
- sensitivity analysis

**behavioural validity**

- model inter-comparisons
- diagnostic indicators
- historical trends
- generalisable historical patterns

- historical simulations
“Climate models ... reproduce many important aspects of observed climate ...” [IPCC AR1 - AR5]

- many simulated (un-tuned) quantities for different processes & scales
- statistical measures of performance

Flato et al. 2013.
Historical simulations are not a common feature of IAM evaluation

- many simulated (un-tuned) quantities for different processes & scales
- statistical measures of performance
- no long-run simulations of aggregate system variables
4.3. How well are we capturing variability across states?

In this section we focus on comparing residential sector projections at the state level. We focus on 2005 for which we have both reported fuel-consumption at the state level and service-level estimates developed for GCAM calibration. Energy consumption by service from RECS is only available for major geographical regions, e.g., New England, West North Central, Pacific, etc. as well as for the four most populous states—New York, Florida, Texas, and California. State-level estimates for 1990 and 2005 were originally created for purposes of GCAM model calibration by downscaling the regional RECS data to the state level on the basis of population, GDP, state-level fuel consumption, and population weighted heating and cooling degree-days. We are, therefore, comparing a model projection with state-level estimates that, in many cases, are produced using the same assumptions as used for the modeling. This is unavoidable, but may overestimate the level of agreement.

As was the case with national level results, the projection for 2005 tends to underestimate electricity consumption and overestimate gas consumption (Fig. 7). The electricity consumption underestimate is fairly consistent across states, with the projection showing an overestimate of larger than 5% for only three states. The largest absolute underestimate for electricity is for states that also have high electricity consumption: California, Florida, and Texas. These are also states with a high number of cooling degree days.

Historical simulations are not a common feature of IAM evaluation

- many simulated (un-tuned) quantities for different processes & scales
- statistical measures of performance

- very limited in scope (process, time horizon)
- divergence -> Δ parameter
4.3. How well are we capturing variability across states?

In this section we focus on comparing residential sector projections at the state level. We focus on 2005 for which we have both reported fuel-consumption at the state level as well as service-level estimates developed for GCAM calibration. Energy consumption by service from RECS is only available for major geographical regions, e.g., New England, West North Central, Pacific, etc. as well as for the four most populous states—New York, Florida, Texas, and California. State-level estimates for 1990 and 2005 were originally created for purposes of GCAM model calibration by downscaling the regional RECS data to the state level on the basis of population, GDP, state-level fuel consumption, and population weighted heating and cooling degree-days. We are, therefore, comparing a model projection with state-level estimates that, in many cases, are produced using the same assumptions as used for the modeling. This is unavoidable, but may overestimate the level of agreement.

As was the case with national level results, the projection for 2005 tends to underestimate electricity consumption and overestimate gas consumption (Fig. 7). The electricity consumption underestimate is fairly consistent across states, with the projection showing an overestimate of larger than 5% for only three states. The largest absolute underestimate for electricity is for states that also have high electricity consumption: California, Florida, and Texas. These are also states with a high number of cooling degree days.

Fig. 4. Historical and projected final energy by fuel for residential sector.

Fig. 5. Heating and cooling energy per unit floorspace (FS) normalized for degree days.

Historical simulations are not a common feature of IAM evaluation

- many simulated (un-tuned) quantities for different processes & scales
- statistical measures of performance
- very limited in scope (process, time horizon)
- divergence -> ∆ parameter

similar issues with behavioural validity testing

over-tuning, non-uniqueness

limited historical conditions
Historical simulations are not a common feature of IAM evaluation

- very limited in scope (process, time horizon)
- divergence -> ∆ parameter

**lack of good observational data**
**heterogeneous causal processes**
*(normative design)*

---

Chaturvedi et al. 2013 | van Ruijven et al. 2011 | Schwanitz 2013 | van Vuuren et al. 2010
The historical record can be used for IAM evaluation in other ways - generalisable historical patterns

- useful learning exercise ...
  - but what then?
- no clear methodology or metrics

- very limited in scope (process, time horizon)
- divergence -> Δ parameter
Simple models help understand representations of key processes embedded in more complex models.

"A complex model may be more realistic, yet ... as we add more factors to a model, the certainty of its predictions may decrease even as our intuitive faith in the model increases."

- elegance vs. elaboration
- simpler models preserved in a ‘hierarchy of models’
Simple models help understand representations of key processes ... but are not common in IAMs

- elegance vs. elaboration
- simpler models preserved in a ‘hierarchy of models’
- ‘SIMPLE’ global agriculture model - biophysical, economic
- historical simulations (1961-2000)
Model inter-comparison projects explore structural uncertainty (across different model representations)

GCMs e.g., CMIP5

- harmonised experiments & results
- model performance metrics
Model inter-comparison is a long tradition for IAMs (9 MIPs contributed 95% of AR5 mitigation scenarios)

GCMs e.g., CMIP5

Snow cover extent change

- harmonised experiments & results
- model performance metrics

IAMs

$\$50 carbon tax (2010), increasing 4% per year—World

- emphasis on robust results
- diagnostic indicators (recent)
- link structure <-> behaviour
Evaluation research for GCMs is generally more developed and prominently reported than for IAMs

<table>
<thead>
<tr>
<th>evaluation method</th>
<th>GCMs vs. IAMs</th>
<th>GCM vs. IAM differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>historical simulations</td>
<td></td>
<td>(1) modelled system</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- underlying principles</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- observational data</td>
</tr>
<tr>
<td>generalisable historical patterns</td>
<td></td>
<td>(2) domain of application</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- uniqueness of insights</td>
</tr>
<tr>
<td>simple models</td>
<td></td>
<td>- expertise of policy users</td>
</tr>
<tr>
<td>model inter-comparisons</td>
<td>+ sensitivity analysis + expert review + documentation ...</td>
<td></td>
</tr>
</tbody>
</table>
Each evaluation method has characteristic strengths and weaknesses

<table>
<thead>
<tr>
<th>evaluation method</th>
<th>strengths</th>
<th>weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>historical simulations</td>
<td>e.g., use of observations</td>
<td>e.g., limited applicability</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(time horizon, processes)</td>
</tr>
<tr>
<td>generalisable historical patterns</td>
<td>e.g., use of observed dynamics</td>
<td>e.g., unclear implications for structural validity</td>
</tr>
<tr>
<td>simple models</td>
<td>e.g., insights robust to structural uncertainty</td>
<td>e.g., attribution of divergence to model differences</td>
</tr>
<tr>
<td>model inter-comparisons</td>
<td>e.g., understanding of key system processes</td>
<td>e.g., lack of realism</td>
</tr>
</tbody>
</table>

+ sensitivity analysis  
+ expert review  
+ documentation
Each evaluation method has **strengths & weaknesses** ... and contributes more to certain **evaluation criteria**

<table>
<thead>
<tr>
<th>evaluation method</th>
<th>appropriateness</th>
<th>interpretability</th>
<th>verifiability</th>
<th>credibility</th>
<th>usefulness</th>
</tr>
</thead>
<tbody>
<tr>
<td>historical simulations</td>
<td>~</td>
<td>~</td>
<td></td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>generalisable historical patterns</td>
<td></td>
<td>~</td>
<td></td>
<td>~</td>
<td></td>
</tr>
<tr>
<td>simple models</td>
<td>✔</td>
<td>✔</td>
<td>~</td>
<td>~</td>
<td></td>
</tr>
<tr>
<td>model inter-comparisons</td>
<td></td>
<td>✔</td>
<td></td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>

+ sensitivity analysis
+ expert review
+ documentation ...
Conclusion: Systematic & more prominent evaluation effort to strengthen and maintain confidence in IAMs

• systematic: multiple methods concurrently

• prominent: concerted, synthesis products

• learning: insights from GCMs

ongoing articulation of the grounds on which IAMs can be declared good enough for their intended uses
Evaluating Model Analysis of Climate Change Mitigation

Charlie Wilson

November 2015, IAMC Conference

(Elmar Kriegler, Jana Schwanitz, Celine Guivarch, Volker Krey, Keywan Riahi, Detlef van Vuuren)