POTSDAM INSTITUTE FOR
CLIMATE IMPACT RESEARCH

New Scientific Working Group (SWG) on Model Evaluation and Diagnostics

Jae Edmonds, Elmar Kriegler, John Weyant

IAMC Annual Meeting 2013

NIES, Tsukuba, 28.11.2013

Leibniz
Gemeinschaft

Why do we need model evaluation & diagnostics?

- **Need to express more clearly how our models relate to reality**
- **Need to better understand why model results differ**
- **Our models are used for policy advice. What is the meaning of the model results for policy?**
- **Policy makers ask for building confidence in model results. The more the model results become relevant, the more confidence building is needed.**
- **There has been a lack of emphasis on model evaluation vs. policy application in the community**
  - ➔ How much time on this vs. development, calibration, and application?
  - ➔ How does this compare to other communities, e.g. climate modeling

Why use models?

- How would maps look like without *cartographers*? *Scientists* can play the role of cartographers for the exploration of the solution map.
- And would maps be of any use without *navigators*? *Policy makers* navigate through the maze of possible solutions in the solution map.

What work has happened recently on the topic?

- **Work on model diagnostics and model validation in PIAMDDI and AMPERE**
- **Work on model documentation in MIPs**
- **Work on hindcasting by the GCAM team**
- **May 2012 – PIAMDDI Workshop, Stanford**
- **November 2012 – Session at 2012 IAMC Meeting, Utrecht**
- **May 2013 – AMPERE Workshop on Model Validation, Seville**
- **Thereafter, establishment of SWG (Chairs: Jae Edmonds, Elmar Kriegler, John Weyant)**
- **November 2013 – here we are in Tsukuba ...**

Some insights from the workshops

Yarman Barlas (Bogazici U):

- Behavioral validity: Matching observations of modelled system
- Structural validity: Not only observations should be matched, but matched for the right reason.

➔ Structural validity is the appropriate category for dynamic system models, including IAMs

➔ Structural validity can not be proven, it is „build up“ in a continuous process of evaluation

Barlas, Y., 1996. Formal aspects of model validity and validation in system dynamics. System Dynamics Review 12, 183–210.
Barlas, Y., Carpenter, S., 1990. Philosophical roots of model validation: Two paradigms. System Dynamics Review 6, 148–166.

Some insights from the workshops

Rob Sargent (Syracuse U):

- Develop models from simple to complex to enable validation (tension with state of play in IAM community)
- Good to have a group on validation. Consider independent verification and validation panels.
- Terminology: Behavior graphs; Stylized facts = observed system behavior; Hindcasting = historical data validation

Sargent, R. G. 2013. "Verification and Validation of Simulation Models", Journal of Simulation 7: 12-24

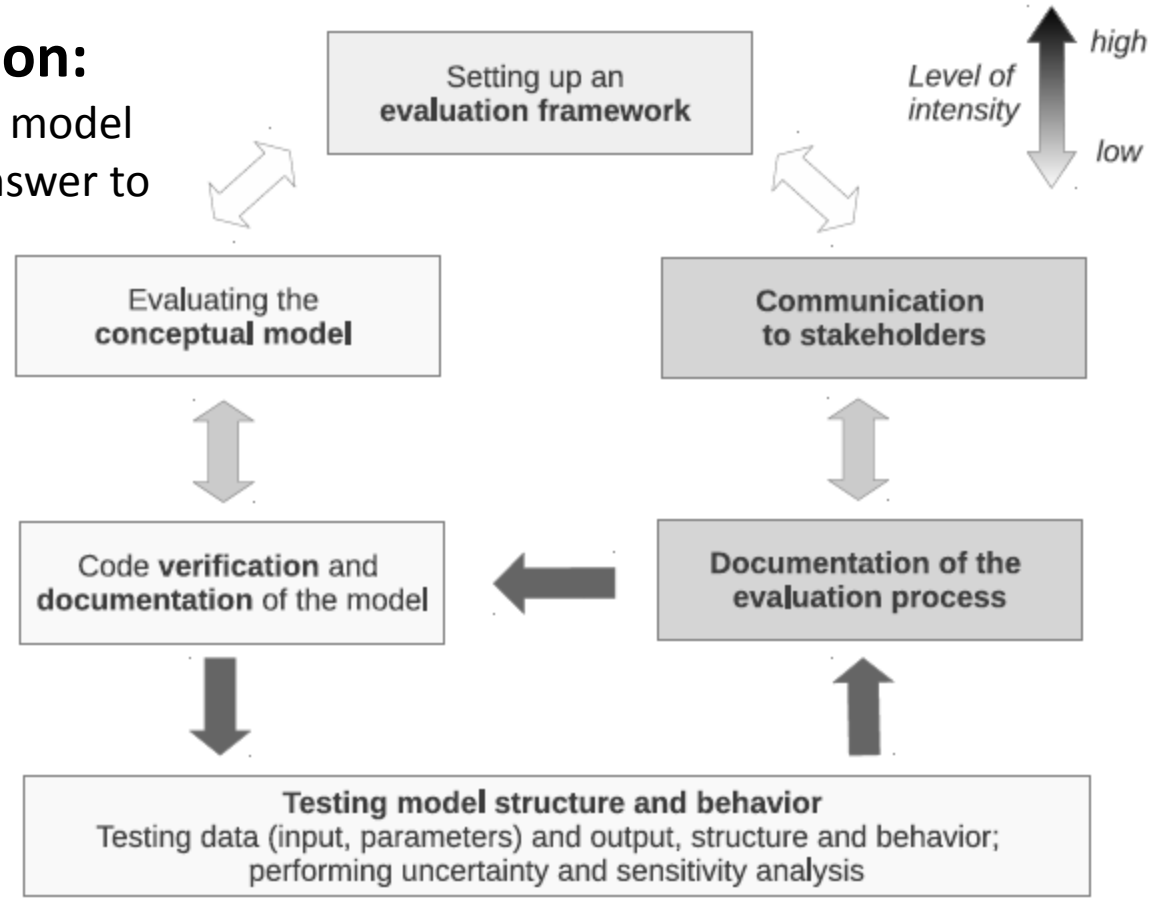
Ben Santer (LBNL):

- Many analogies to diagnostics and evaluation work in climate modeling community

Conceptual approach to validation of IAMs

Key evaluation question:

Can we confidently apply the model to deliver a well-grounded answer to the group of users?



J. Schwanitz (2013) Evaluating integrated assessment models of global climate change. Environmental Modelling & Software

Two questions that the Model Evaluation and Diagnostics SWG is presently addressing

- **What activities are within the scope of the Model Evaluation and Diagnostics SWG?**
  - Model diagnostics (sensitivity analysis)
  - Hind Casting
  - Stylized Facts (= Historic behavior patterns)
  - Model documentation
    (together with Data Management SWG)
  - *Uncertainty Analysis (?)*

- **What activities should the SWG undertake?**

Model Diagnostics

- *Purpose: to provide a measure to easily understand differences across studies as a result of different models.*

- Development of a set of routinely calculated Indicators (comparable to climate sensitivity for ESM/GCM models)
  - Choice of specific indicator variables, e.g. elasticity of CO_2 emissions for a given carbon price.
  - Need to identify indicators based on the identification of key questions the models should be able to address
  - Classifies models as sensitive / insensitive without providing a clear explanation why a model is sensitive or what would be a good value

Model Evaluation

- **Stylized facts**: comparing scenarios of future events with historical experiences

- **Hind casting**: comparing hind-cast scenarios to history.

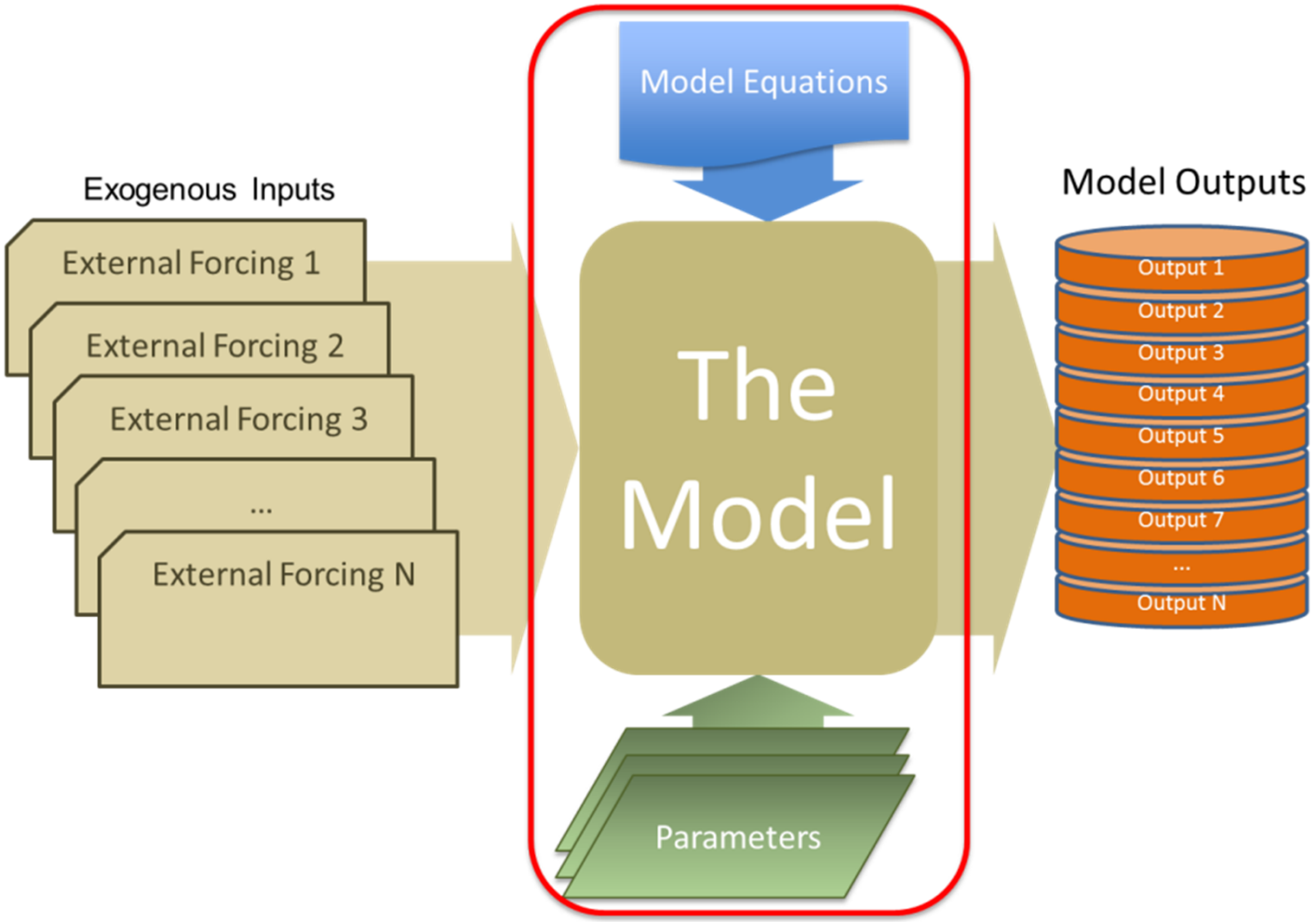
This is a relatively new area for the IAM community

Motivation for Hind Casting

- Integrated assessment modelers have and will continue to be asked by potential users and critics:
  - "How do we know that if we gave you all of the right information about future states of the world, that your model would give us an accurate prediction of future model outputs?"
  - "Have you ever started from a historical year and predicted the present?"

- We also get accused of "just making our stuff up."

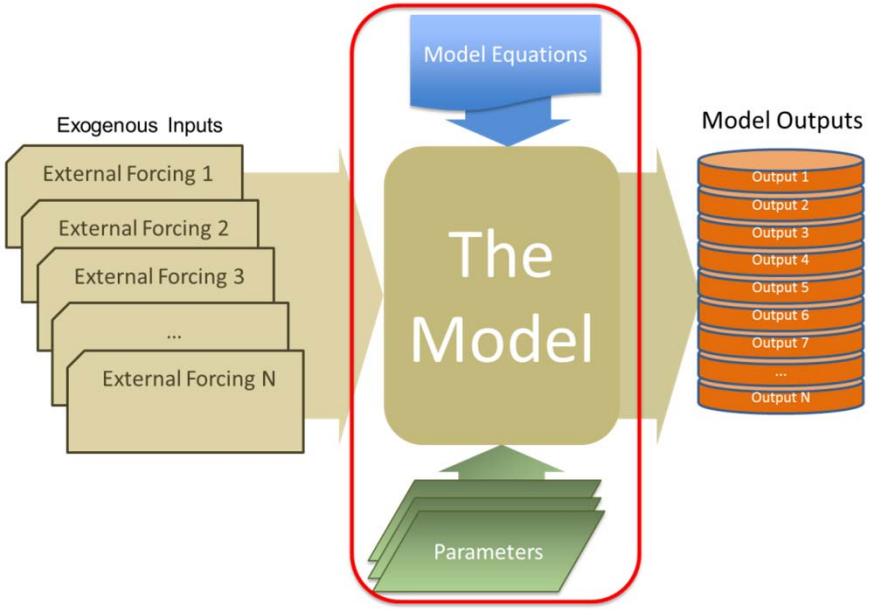
Contingent not absolute prediction



Is it fair to ask IAMs to predict the future?

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

- We expect weathermen, macro-economists and political scientists to make predictions all the time and compare their forecasts to observations.
- Even if we are primarily about insights, those insights need to be developed in the context of a system of analysis that, if given accurate exogenous variable values, would generate an accurate representation of real events.
- Otherwise we don't have insights.



What do we need to get started in the business of hind-casting?

- **Historical Data**
  - We need an agreed upon history (not to be confused with what really happened)
  - Data to initialize the models in a prior period—all of the data for our models which are, by nature extensive (energy, economy, land use, land cover, carbon stocks, capital vintages)
  - Data to describe events in the periods between the initialization data and the present

- **Method**
  - What questions do we want our models to answer?
  - What are we testing?
  - How are we going to test our model output?
  - What will our performance measure be?

Just setting up hindcasting experiments will help clarify many of our critical issues.

Past Performance is No Guarantee of Future Results

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

- Getting it wrong can be more instructive than getting it right
  - Understanding why a hind cast failed will yield important insights.
  - Understanding why a hind cast failed will help point the way toward model improvements.

- Getting the hind cast right, does not guarantee that future events will be predicted accurately.
  - Much of the uncertainty about the future is embedded in uncertainty surrounding future forcing variables, e.g. population, GDP, technology, and policy.
  - Many of the scenarios that are routinely examined take the model outside of the bounds of past experience.
    - E.g. Carbon taxes push energy prices outside the range of historical experiences.
    - E.g. reference scenarios take developed economies outside the range of per capita income found in the historical record.

What can the IAMC SWG on Model Evaluation and Diagnostics do?

- Keep track of organized activities, e.g. ADVANCE, PIAMDDI, PNNL, other.
- Provide a forum for coordination:
  - Provide a place where individual projects can coordinate the design of their activities.
  - Avoid reinventing the wheel
  - Share data and methods
- Next step: Establish community standards
  - Diagnostic indicators and experiments
  - Hindcasting Experiments
  - Behavioral patterns with explanatory power for IAMs

Questions?

Example of Validation and Diagnostics Work in AMPERE

Elmar Kriegler, Jana Schwanitz

IAMC Annual Meeting 2013

NIES, Tsukuba, 28.11.2013

PIK

SEVENTH FRAMEWORK PROGRAMME

Model diagnostics – Motivation

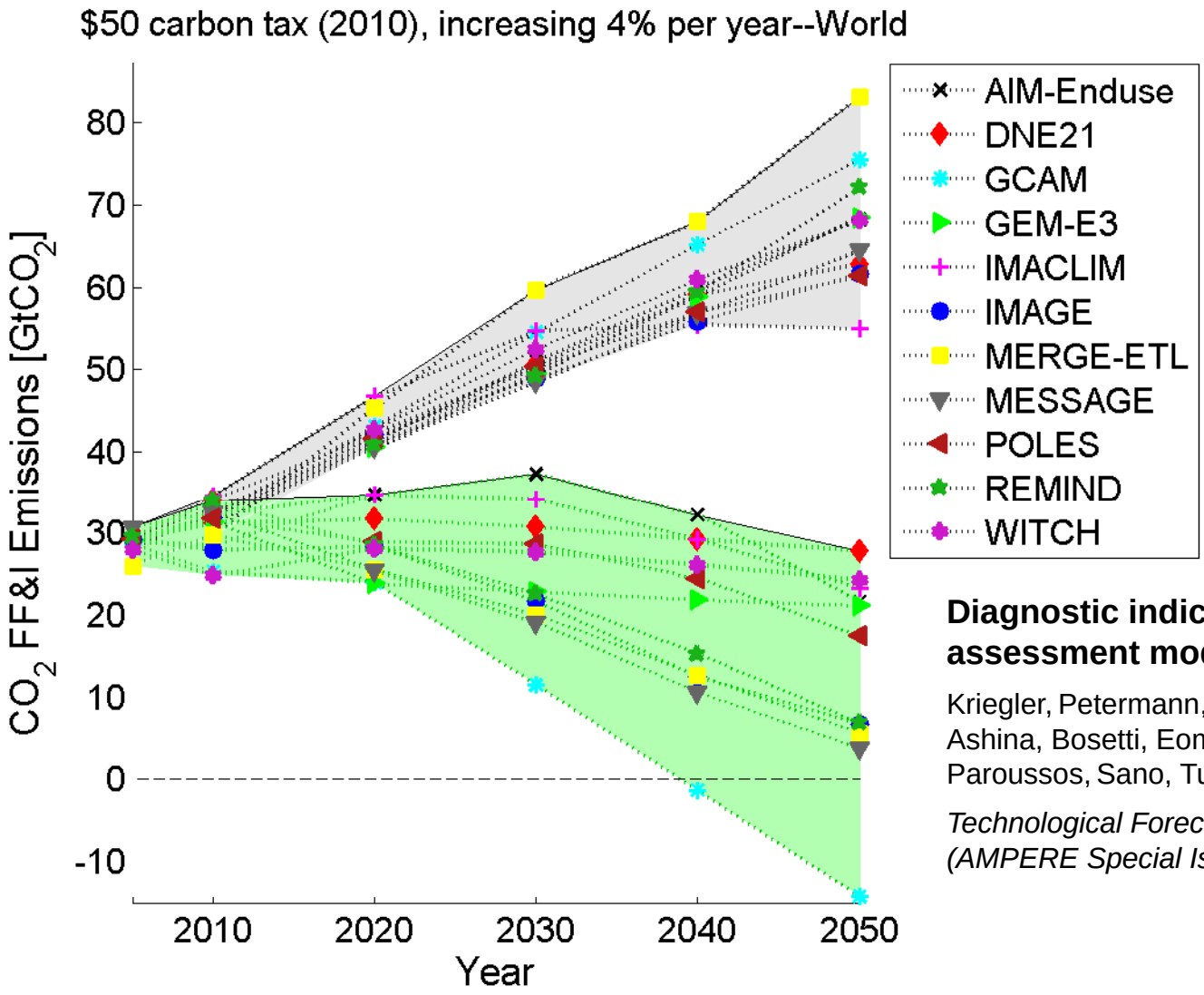
Expectation of funding institution (EC DG Research):

- Improve knowledge on climate change mitigation costs
- Provide operational information on the interpretation of the model outputs and uncertainties
- Increased consistency in cost-related information for policy making

Goal of diagnostic work in AMPERE:

- Identify indicators of model behaviour that help to explain the spread of model results in key quantities (e.g. mitigation costs, decarbonization rates)
- Develop rough model classification scheme that can assist the comparative analysis of model results

Experiment: Emissions response to carbon tax



Diagnostic indicators for integrated assessment models of climate policy

Kriegler, Petermann, Krey, Schwanitz, Luderer, Ashina, Bosetti, Eom, Kitous, Méjean, Paroussos, Sano, Turton, Wilson, Van Vuuren

Technological Forecasting and Social Change (AMPERE Special Issue), forthcoming

Selection criteria for diagnostic indicators

- identification of heterogeneity in model responses
- relevance for climate policy analysis
- applicability to diverse models
- accessibility and ease of use

Model	Relative Abatement Index	CoEI Indicator	Transformation Index (primary energy)	Cost per Abatement Value	Model type	Classification
…	…	…	…	…	PE or GE	…

Characterize system response to emissions price **X** **Characterizes cost response to emissions price** **= Magnitude of mitigation costs**

➔ Low system response leads to high carbon price for fixed emissions reduction

Model classification („fingerprints“)

Model	Relative Abatement Index	CoEI Indicator	Transformation Index (primary energy)	Cost per Abatement Value	Classification
AIM-Enduse	Low	Mixed	Mixed	TBD	PE – med response
DNE21+	Low	High	Low	Mixed	PE – low response
GCAM		Low	High	Medium	PE – high response
GEM-E3		High	TBD	Medium	GE – low response
IMACLIM	Low	High	Mixed	High	GE – low response
IMAGE	High	Low	Mixed	Low	PE – high response
MERGE-ETL		Low	High	Low	GE – high response
MESSAGE		Low	High	Low	GE – high response
POLES	Mix	Mixed	Low	Low	PE – med response
REMIND	High	Low	High	Medium	GE – high response
WITCH	Low	High	Low	Medium	GE – low response

Highest cost

Lowest cost

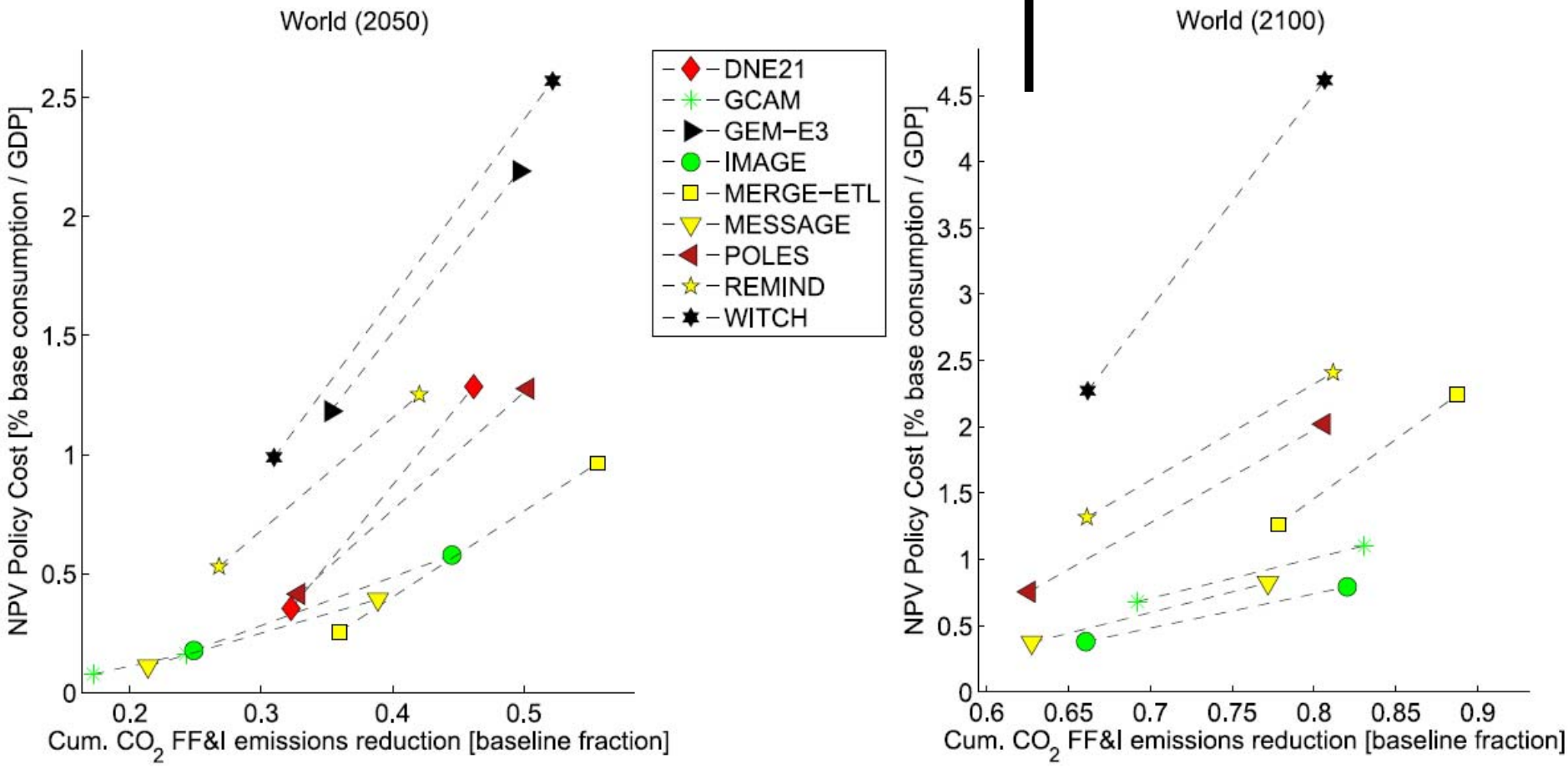
Kriegler et al. (2013) Tech. For. & Soc. Change, forthcoming

SEVENTH FRAMEWORK PROGRAMME AMPERE

Mitigation costs in AMPERE WP3 study

(450/ 550 ppm CO2e)

IMACLIM: 7% (550 ppm),
18% (450 ppm) for 2010-2100



Validation using stylized facts

Ongoing work by Jana Schwanitz et al. in AMPERE: Evaluating integrated assessment models with stylized facts - a multimodel analysis

Objective: Systematic evaluation of IAMs with stylized facts. Long term goal could be a community-wide list of stylized facts for this purpose.

Criteria for selection:

- Evaluation value

  acceptance, relevance, endogenous and exogenous model results, transperancy

- Completeness

  capturing important system processes and scales

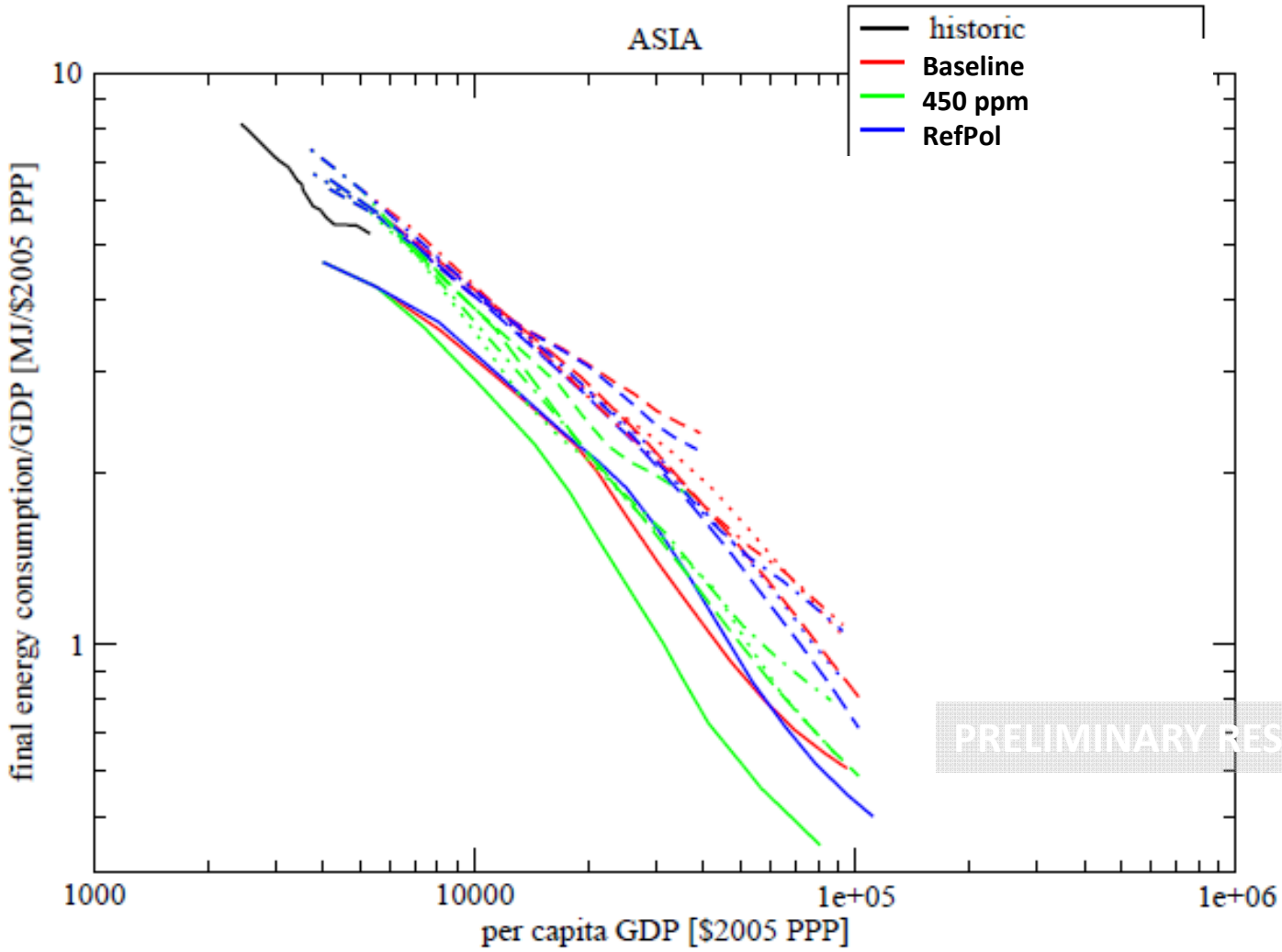
- Broad applicability

Validation using stylized facts

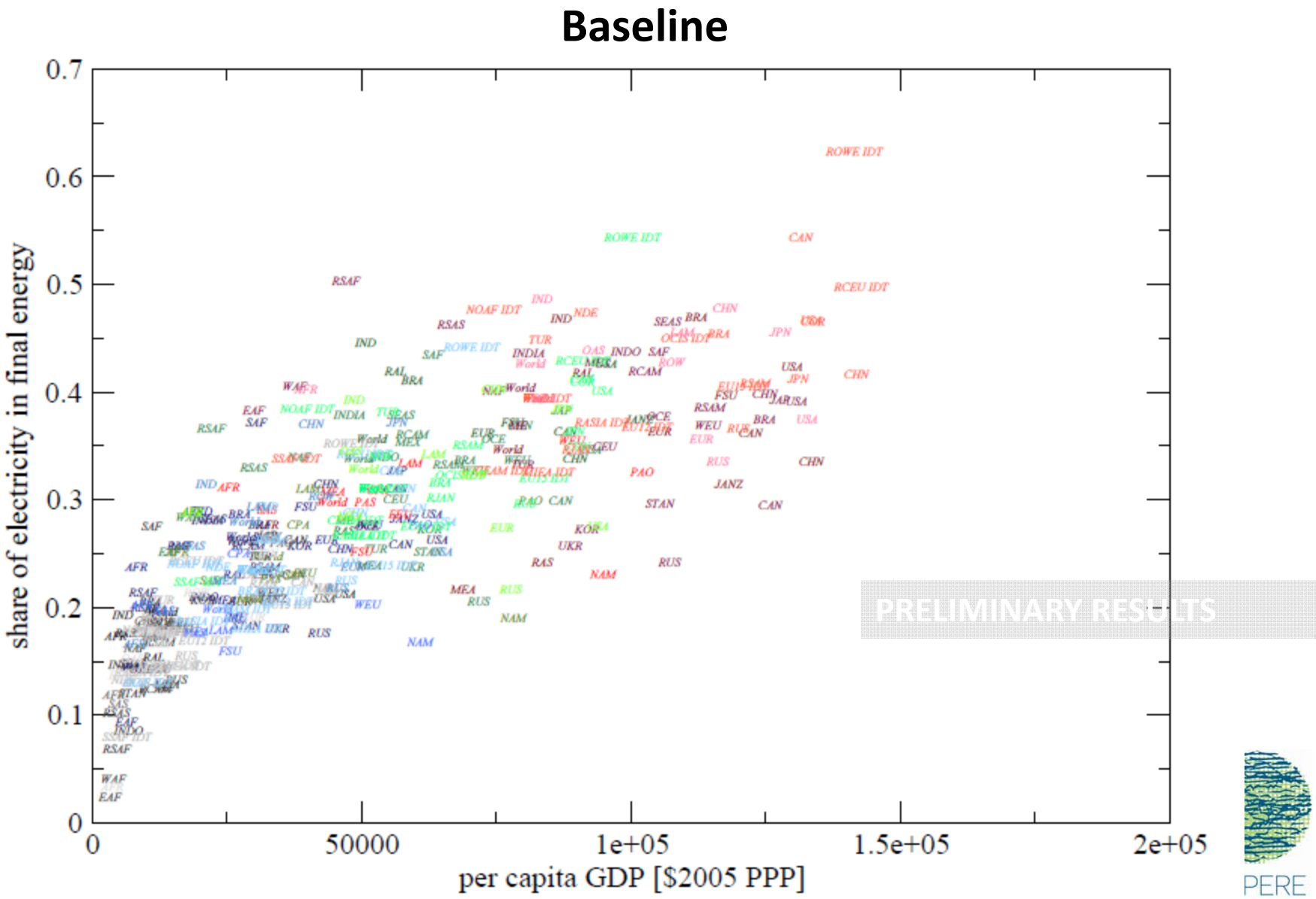
Relevant stylized facts relating to the energy transition
(preliminary list; comments, suggestions welcome)

Stylized Fact	Expectation
(Log-log) relationship between final energy intensity (of GDP) and GDP per capita	Holds in all scenarios
PE (FE?) per capita increases with per capita income	Holds in baseline scenarios. Saturates in mitigation scenarios?
Electricity share in FE increases & solids share in FE decreases with per capita income	Holds in all scenarios. Acceleration in mitigation scenarios
U-shape of industry share in FE with increasing per capita income	Holds in all scenarios
Increasing share of services/transport in FE with increasing per capita income	Saturation in the long term? Earlier in mitigation scenarios

Example: FE / cap vs GDP / cap



Example: Electricity share in FE vs. GDP / cap



Community adoption?

Once work on hindcasting, diagnostics, and evaluation using stylized facts **is mature enough**, community could take up standards as part of the evaluation process

- Hindcasting: Design of experiment(s) and establishment of historic dataset(s)
- Diagnostics: Definition of indicators and standard experiments to derive them (could be semi-automatized ➔ ADVANCE)
- Stylized facts: Identification of robust and relevant stylized facts

Standardization could be task of the Evaluation & Diagnostics SWG in the longer term.