FACCE-MACSUR

Protocol for model evaluation

Gianni Bellocchi[a,*], Mike Rivington[b], Marco Acutis[c]

[a] Grassland Ecosystem Research Unit, French National Institute for Agricultural Research, 5 Chemin de Beaulieu, 63039 Clermont-Ferrand, France
[b] The James Hutton Institute, Craigiebuckler, Aberdeen, AB15 8QH, United Kingdom
[c] Department of Agricultural and Environmental Sciences - Production, Landscape, Agroenergy, University of Milan, Italy

*gianni.bellocchi@clermont.inra.fr

Instrument:	Joint Programming Initiative
Topic:	Agriculture, Food Security, and Climate Change
Project:	Modelling European Agriculture with Climate Change for Food Security (FACCE-MACSUR)
Start date of project:	1 June 2012
Duration:	36 months
Theme, Work Package:	LiveM 2
Deliverable reference num.:	D-L2.2
Deliverable lead partner:	INRA
Due date of deliverable:	month 24
Submission date:	2014-06-11
Confidential till:	–

Revision	Changes	Date
1.0	First Release	2014-06-11

Abstract/Executive summary

This deliverable focuses on the development of methods for model evaluation in order to have unambiguous indications derived from the use of several evaluation metrics. The information about model quality is aggregated into a single indicator using a fuzzy expert system that can be applied to a wide range of model estimates where suitable test data are available. This is a cross-cutting activity between CropM (C1.4) and LiveM (L2.2).

Table of Contents

Introduction

This protocol is the first major attempt to lay the groundwork for good practice standards of model evaluation based on the use of modern concepts and criteria. The basis of the protocol stems from the progresses made over the last two decades in setting new horizons for model performance and on the problematic interpretations made of model evaluation.

The main important progresses made in the domain can be summarized as follows:

- aggregation of multiple evaluation metrics into integrated indicators (based on the fuzzy logic principle, after Bellocchi et al., 2002a)
- assessment of model departure from observations with respect to an external variable (pattern indices by Donatelli et al., 2004)
- inclusion, in the evaluation of models, of other measures than performance metrics, such as sensitivity analysis measures and information criteria for model selection (Confalonieri et al., 2009a), and consideration by expert stakeholders (Alexandrov et al., 2011)
- elaboration of the model robustness concept (Confalonieri et al., 2010a)
- elaboration of the model plasticity concept (Confalonieri et al., 2012)

Such evolution in model evaluation, yet accompanied by the creation of dedicated software tools (Fila et al., 2003a, b; Criscuolo et al., 2005; Tedeschi, 2006; Olesen and Chang, 2010), has recently culminated in a review article (Bellocchi et al., 2010) as well as position papers (Alexandrov et al., 2011; Bennett et al., 2013) of the International Environmental Modelling & Software Society (http://www.iemss.org) with the aim of characterising the performance of models and providing standards for publishing models in forms suitable for use by broad communities (Laniak et al., 2013). Alternative validation strategies were documented by Richter et al. (2012) and Ritter et al. (2013). Some novel ideas about model evaluation have also found application for validating analytical methods (e.g. Acutis et al., 2007; Bellocchi et al., 2008) to complement standard assessment approaches of the International Organization for Standardization (http://www.iso.org).

Also graphical tools have been developed to help assessing the quality of model performances (e.g. Taylor diagrams, Taylor, 2001).

General goals

The primary goal is to evaluate the quality of crop and grassland models in predicting production and other variables while considering integrated multiple metrics in order to have unambiguous indications about model accuracy and robustness under a variety of conditions. Accurate and robust models offer reduced uncertainties under scenarios where no calibration data exist (e.g. climate scenarios and areas not covered by experimental sites). These goals are about evaluation of models under conditions of 'known unknowns' such as models which could not represent things like pests, diseases or physical damage.

Objectives

1. To evaluate crop and grassland models in response to climatic and management factors by comparing the simulated results with the observed data by:

a. Identifying common sets of input and outputs (mainly production outputs);
b. Identifying common evaluation metrics;
c. Identifying ranges of acceptability and relative weights for each metric.

2. To document model evaluation experiences against test cases and assess them with respect to alternative models (e.g. comparing the results obtained with different models for the same crop, cropping system or grassland).
3. To formulate standard criteria for model evaluation (and creation of exemplary evaluation tools).
4) To expand the concept of robustness in the use of crop/grassland models to simulate yield or other variables of interest under climate change conditions, towards including a variety of meteorological and soil conditions (with respect to the original formulation of the robustness index based on ET_0 and precipitation, other variables such as temperature and soil properties should be included).

Evaluation strategy

Model evaluation cannot be performed in an absolute way looking at one (or few) metrics (indices or test statistics) for summarizing some model behaviors. For a long time authors have looked at the evaluation problem as if it was mainly an issue of selecting some appropriate evaluation metrics and assessing their values (e.g. Nash and Sutcliffe, 1970; Willmott, 1981; Greenwood et al., 1985; Loague and Green, 1991; Stöckle et al., 2004). Also recently, authors have been working towards the further development of evaluation metrics (e.g. Jain and Sudheer, 2008; Willmott et al., 2012; Legates and McCabe, 2013). However, it has become clear since long ago that each kind of problem faced with modelling tools through simulation processes needs a specific evaluation scheme (e.g. Bellocchi et al., 2002a for evaluation of solar radiation models; Confalonieri et al., 2006 for comparison of rice growth and yield models; Moriasi et al., 2007 for evaluation of watershed runoff estimations; Bregaglio et al., 2010, 2011 for simulation of relative humidity and leaf wetness, respectively). The lack of precise and undisputable criteria to consider a specific metric as more effective than others, and the multiplicity of aspects to be accounted for a multi-perspective evaluation of model performance, logically leads to some use of composite metrics for model validation (e.g. Bellocchi et al., 2002b; Diodato et al., 2007a, b; Rivington et al., 2007; Confalonieri et al., 2009b, 2010b). With a composite method, the best is obtained with combining the metrics, while also having the information provided by the individual metrics. In such respect, composition of metrics is a shift of paradigm from merely selecting the best out of a set of evaluation metrics.
A problem only partially faced by the actual available knowledge on model evaluation is how to handle multiple outcomes from models. Virtually all cropping system and grassland models offer several relevant outputs such as yield, nitrogen concentration in soil layers, nitrogen and pesticides leaching, water runoff, soil erosion, evolution in time of soil organic matter, etc. These outputs are produced at different space and time scales ranging from daily (or sub-daily) to yearly outputs and from soil layer to site, catchment or region. To reduce the user effort, a modular model allows simulating each process according to a modelling solution that the user may select out of alternate solutions based on his/her knowledge of the system, data availability, computing resources, etc. (Donatelli and Rizzoli, 2008). There is thus the need to understand if a single model can cover all the required outputs simultaneously, offering an implicit warranty of coherence of all aspects of the simulation, or several models need to be used. For crop models, the need of simultaneously evaluating several outputs was highlighted by Wallach (2006). An attempt to address the same scenario with hydrological models was done by Confalonieri et al. (2010b) by using fuzzy-logic based rules. In principle, fuzzy logic offers again a way to aggregate several metrics in a few or in one indicator. Here, the risk is either to create an excessively complex evaluation scheme, or a too simple one, thus reducing the problem of multiple output evaluation to a weighted sum of performance metrics.
There is a challenge to develop disciplined answers to the issues in the debate opened by Matthews et al. (2011) targeting at shifting towards model “outcome” rather than merely

model “output” assessment. An interesting way to develop a fuzzy-logic based scheme is to explicitly involve a network of experts in a participatory activity through group discussions and interviews. In this cross-cutting activity we are proposing a three-step approach for the definition of a procedure for model evaluation:

i) collection of a large number of model evaluation metrics, including their characteristics, pros and cons,

ii) definition of the minimum dataset (MDS) of metrics needed for model evaluation on the basis of expert’s opinions and their factual grounding, and

iii) fuzzy-based aggregation of the variables belonging to the MDS.

The procedure will be tested on simulation results of the models used in the inter-comparison tasks (C1.5, L2.4), submitted to the expert panel, and possibly adjusted to expand consensus by enhancing dialogue and joint efforts.

An example of this approach is given in Carozzi et al. (2013) to assess soil quality under different options for soil management.

Basic components for model evaluation

The multi-metric, fuzzy-logic based approach adopted by Confalonieri et al. (2009a) is the basis for model assessment in a comparative fashion (Figure 1).

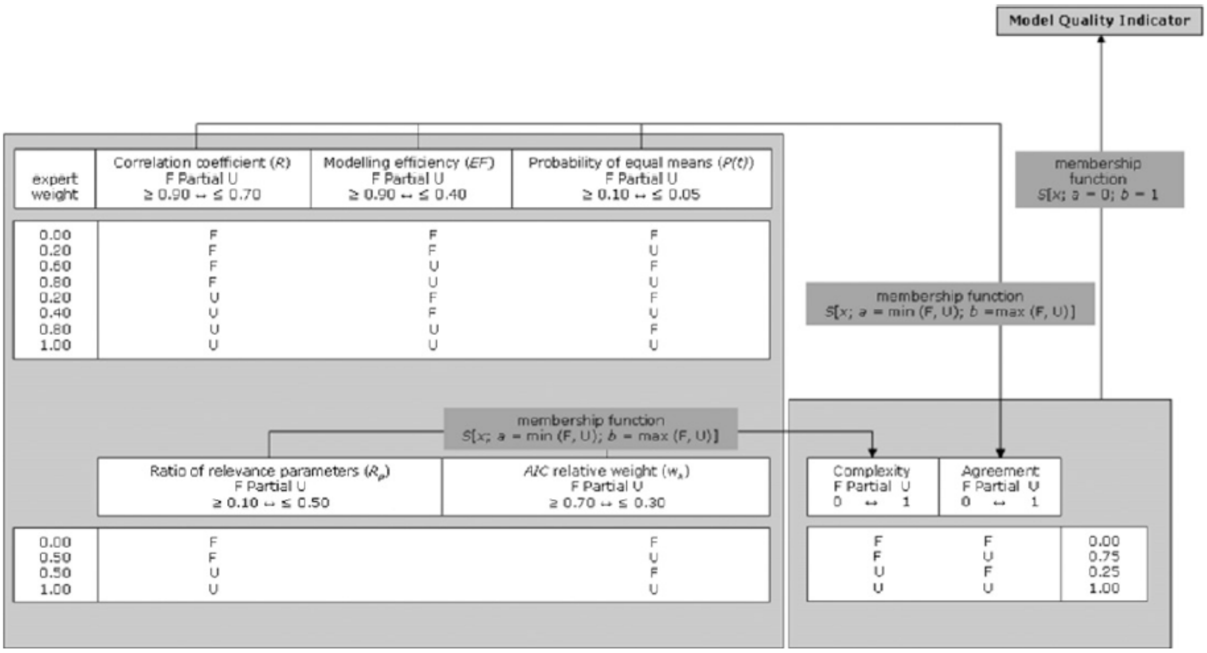


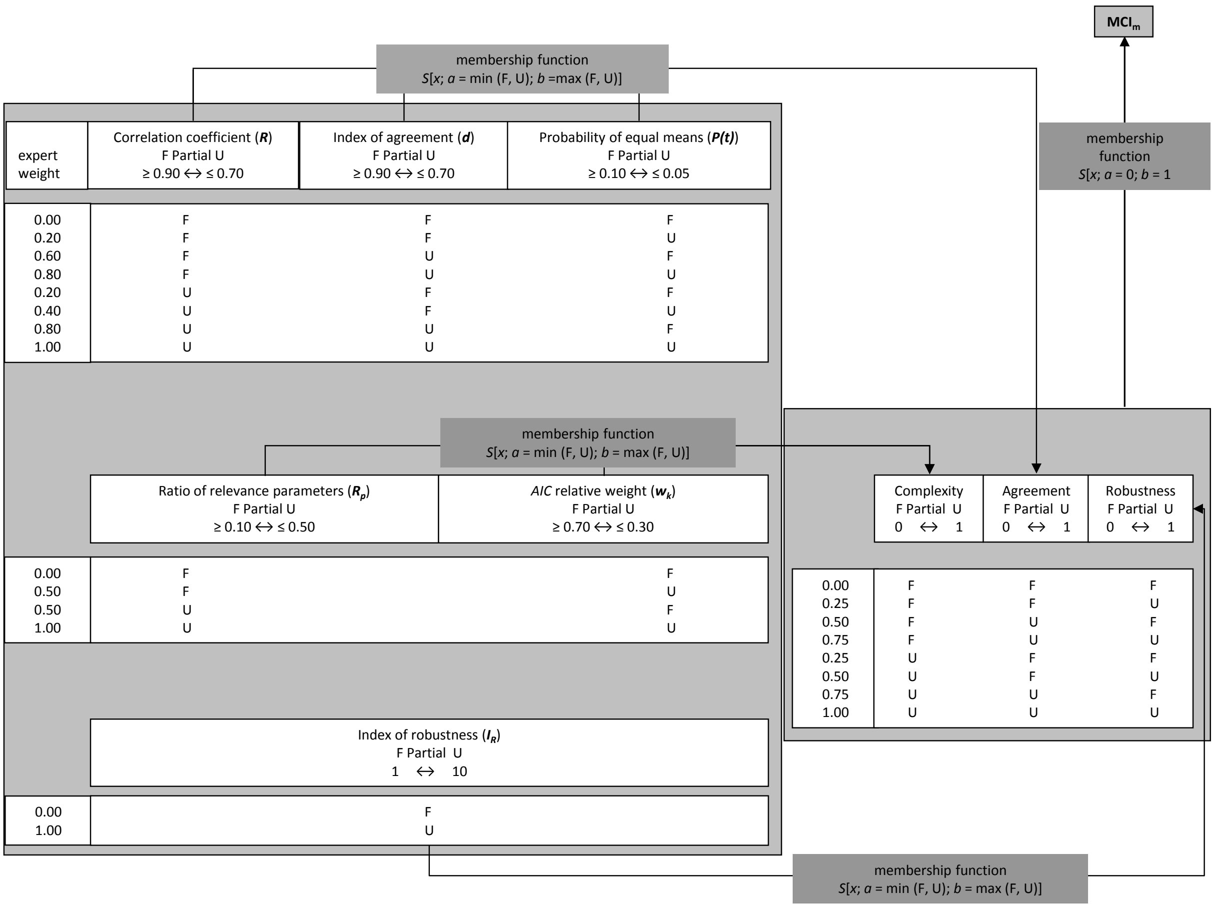
Figure 1. Structure of the Model Quality Indicator (MQI) assessment method, where: *EF*, modelling efficiency; *P*(*t*), Student *t*-test probability of null mean difference between predictions and observations; *R*, correlation coefficient of predictions versus observations; R_p, ratio of relevant model parameters over total number of parameters; w_k, Akaike Information Criterion (*AIC*) ratio; F, favorable threshold; U, unfavorable threshold; *S*, S-shaped membership function; *x*, value of metric; *a*, minimum value between F and U; *b*, maximum value between F and U. Expert weights are assigned as follows: 0.20, 0.60 and 0.20 to *R*, *EF* and *P(t)* in module Agreement; 0.50 and 0.50 to R_p and w_k in module complexity; 0.25 and 0.75 to Complexity and Agreement in the indicator.

The Model Quality Indicator (MQI) was obtained by combining (via fuzzy-logic based weighting) performance metrics (*R*, correlation coefficient between observations and simulations; *EF*, modelling efficiency; *P(t)*, Student-t test probability of equal means between observation and simulations) as well as components of model structure (relevant over total parameters ratio and Akaike Information Criterion-based indicator of the loss of performance as the number of parameters in the model decreases).

In a model inter-comparison exercise, MQI allows ranking best- to worst-performing models not only at the output level (Agreement) but also regarding the parameterization effort (Complexity).
The originally developed indicator targeted the evaluation of model estimates of above-ground biomass under potential conditions. With the main focus on plant growth and development, options for extending this approach to actual conditions could include:

- The number of sensitive and total parameters of the plant modelling structure under actual conditions
- A model robustness measure in the fuzzy-logic based framework to account for site-to-site differences
- Evaluation of model performance with respect to other output variables than above-ground biomass (e.g. soil water content, carbon fluxes, etc.)

Based on the above items, the following fuzzy-logic based multi-metric evaluation framework is proposed (Figure 2). This is meant for the evaluation of one output variable. In case of multiple outputs, the results obtained by applying the same procedure to each output will be used for further analysis and presentation of results.



expert weight	Correlation coefficient (R)	Index of agreement (d)	Probability of equal means ($P(t)$)
0.00	F	F	F
0.20	F	F	U
0.60	F	U	F
0.80	F	U	U
0.20	U	F	F
0.40	U	F	U
0.80	U	U	F
1.00	U	U	U

expert weight	Ratio of relevance parameters (R_p)	AIC relative weight (w_k)
0.00	F	F
0.50	F	U
0.50	U	F
1.00	U	U

expert weight	Index of robustness (I_R)
0.00	F
1.00	U

expert weight	Complexity	Agreement	Robustness
0.00	F	F	F
0.25	F	F	U
0.50	F	U	F
0.75	F	U	U
0.25	U	F	F
0.50	U	F	U
0.75	U	U	F
1.00	U	U	U

Figure 2. Structure of the MQI$_m$ assessment method, where: *d*, index of agreement; *P*(*t*), Student *t*-test probability of null mean difference between predictions and observations; *R*, correlation coefficient of predictions versus observations; R_p, ratio of relevant model parameters over total number of parameters; w_k, Akaike Information Criterion (*AIC*) ratio; I_R, index of robustness (see also Table 1); F, favorable threshold; U, unfavorable threshold; *S*, S-shaped membership function; *x*, value of metric; *a*, minimum value between F and U; *b*, maximum value between F and U. Expert weights are assigned as follows: 0.20, 0.60 and 0.20 to *R*, *d* and *P(t)* in module Agreement; 0.50 and 0.50 to R_p and w_k in module complexity; 0.25, 0.50 and 0.25 to Complexity, Agreement and Robustness in the indicator.

In Figure 2:

- MQI$_m$ stands for Model Quality Indicator for multi-site evaluation
- It is composed of three modules: Agreement, Complexity, Robustness

- The module agreement is made of three basic metrics: Pearson's correlation coefficient (*R*), Willmott's index of agreement (*d*), Student-t probability of equal means for paired data (*P(t)*)
- The module complexity is made of two basic metrics: relevant over total parameters ratio (R_p) and a weighed measure (w_k) of the Akaike's Information Criterion (*AIC*)
- For Agreement and Complexity, basic metrics values are the average of values calculated from the simulations at multiple sites
- The module Robustness is made of one basic metric: index of robustness (I_R)

Single-site evaluation is performed with an indicator, MQI_s, similar to the MQI of Figure 1, in which modelling efficiency (*EF*) is replaced by Willmott's index of agreement (*d*) in module Agreement. MQI_m (Figure 2) is thus an extension of MQI_s to multiple sites. As *EF* is a component of the index of robustness (I_R), duplication was avoided by replacing it by *d*.

Table 1. Multiple-metrics assessment method: modules and basic metrics.

Module	Performance measure	Equation	Unit	Value range and purpose
Agreement	Pearson's correlation coefficient (*R*) between estimates and measurements	$R = \left[\frac{\sum_{i=1}^{n} (P_i - O_i) \cdot (O_i - \overline{O})}{\sqrt{\sum_{i=1}^{n} (P_i - \overline{P})^2 \cdot \sum_{i=1}^{n} (O_i - \overline{O})^2}} \right]^{0.5}$	-	-1 (anti-correlation) to 1 (perfect correlation): the closer the values are to 1, the better performing the model
	d, index of agreement	$d = 1 - \frac{\sum (P_i - O_i)^2}{\sum_{i=1}^{n} (\lvert P_i - \overline{O} \rvert + \lvert O_i - \overline{O} \rvert)^2}$	-	0 (absence of agreement) to 1 (perfect agreement): the closer the values are to 1, the better performing the model
	P(t), Paired Student t-test probability of means being equal	$P(t) = P\left(\frac{\overline{D}}{S_D / \sqrt{n}} \right)$	-	0 (absence of agreement) to 1 (perfect agreement): the closer the values are to 1, the better performing the model
Complexity	R_p, relevant parameter ratio	$R_p = \frac{S}{T}$	-	0 (absence of relevant parameters) to 1 (all parameters are relevant): the closer the values are to 0, the simpler the model use
	w_k, Akaike Information Criterion ratio	$w_k = \frac{e^{-\frac{AIC_k}{2}}}{\sum_{k=1}^{p} e^{-\frac{AIC_k}{2}}}$	-	0 (best model out of a set) to 1 (worst model out of a set): the closer the values are to 0, the simpler and better performing the model
Robustness	I_R, index of robustness	$I_R = \frac{s_{EF}}{s_{SAM}}$	-	0 (perfect robustness) to positive infinity (absence of robustness): the closer the values

				are to 0, the more robust the model
Computational details	$\overline{D}$, average of the differences between *E* predicted and observed values	$\overline{D} = \frac{\sum_{i=1}^{n}(P_i - O_i)}{n}$	Unit of the variable	-
	$\overline{O}$, mean of observed values	$\overline{O} = \frac{\sum_{i=1}^{n} O_i}{n}$	Unit of the variable	-
	s_D, standard deviation of the differences between estimated and observed values	$s_D = \sqrt{\frac{\sum_{i=1}^{n}(D_i - \overline{D})^2}{n-1}}$		
	AIC, Akaike Information Criterion	$AIC = n \cdot log(MSE) + 2 \cdot T$	-	negative infinity (optimum) to positive infinity: the closer the values are to negative infinity, the better performing the model
	EF, modelling efficiency	$\mathrm{EF} = 1 - \frac{\sum_{i=1}^{n}(\mathrm{P_i} - \mathrm{O_i})^2}{\sum_{i=1}^{n}(\mathrm{O_i} - \overline{\mathrm{O}})^2}$	-	negative infinity to 1 (optimum): the closer the values are to 1, the more efficient the model with respect to observed mean
	SAM, standardized agro-meteorological metric	$\mathrm{SAM} = \frac{Rain - ET_0}{Rain + ET_0}$	-	-1 (no rain, water deficit) to 1 (no ET_0, water surplus): the closer the values are to 0, the more balanced the water budget
	s_{EF}, standard deviation of *EF* values	$s_{EF} = \sqrt{\frac{\sum_{j=1}^{s}(EF_j - \overline{EF})^2}{n-1}}$	-	0 (optimum) to positive infinity: the closer the values are to 0, the more robust the model
	s_{SAM}, standard deviation of *SAM* values	$s_{SAM} = \sqrt{\frac{\sum_{j=1}^{s}(SAM_j - \overline{SAM})^2}{n-1}}$	-	0 (optimum) to positive infinity: the closer the values are to 0 the more similar site conditions
	D, difference between predicted and observed values	$D = P_i - O_i$	Unit of the variable	negative infinity (underestimation) to positive infinity (overestimation): the closer the values are to 0, the less biased the model
	S, number of	-	-	0 (optimum) to

	relevant parameters in a model[1]			positive infinity: the closer the values to 0 the easier model parameterization
	T, number of parameters in a model[2]	-	-	0 (optimum) to positive infinity: the closer the values to 0 the simpler the model
	k, each of models being compared	-	-	-
	p, number of models being compared	-	-	-
	m, number of sites being simulated			
	j, each of sites being simulated			
	P, predicted value	-	Unit of the variable	-
	O, observed value	-	Unit of the variable	-
	n, number of *P*/*O* pairs	-	-	-
	i, each of *P*/*O* pairs	-	-	-

[1] Relevant parameters are those which the model is most sensitive to. They are from formal sensitivity analysis exercises or based the understanding of the modelling context and scope (e.g. the parameters which are more frequently considered for calibration). Depending on the purpose of evaluation, a reduced set of relevant parameters can be built (for instance, only parameters of the plant).

[2] The total number of model parameters is restricted to parameters accessible to users (parameters embedded in the code, but not available to users, are not considered). Depending on the purpose of evaluation, a reduced set of parameters can be built (for instance, only parameters of the plant), and an upper threshold can be set at a level which reflects a high model complexity (for instance, if total parameters is greater than 100, then *T*=100).

Conclusions

The indicator's settings were evaluated via a questionnaire-based survey (Appendix A), whose results are reported in Appendix B. Overall the answers received corroborate the choices made, whereas the approach to robustness requires further assessment. The indicator for model evaluation, facilitated by ready-to-use software (Appendix 3), will be applied to simulation results from CropM and LiveM actions.

Acknowledgements

This report is a contribution to the FACCE MACSUR knowledge hub. We thank all partners who have contributed to answer the questionnaire.

References

Acutis, M., Trevisiol, P., Confalonieri, R., Bellocchi, G., Grazioli, E., van den Eede, G., Paoletti, C., 2007. AMPE: a software tool for analytical method validation. Journal of AOAC International 90, 1432-1438.

Alexandrov, G.A, Ames, D., Bellocchi, G., Bruen, M., Crout, N., Erechtchoukova, M., Hildebrandt, A., Hoffman, F., Jackisch, C., Khaiter, P., Mannina, G., Matsunaga, T., Purucker, S.T., Rivington, M., Samaniego, L., 2011. Technical assessment and evaluation of environmental models and software. Environmental Modelling & Software 26, 328-336.

Bellocchi, G., Acutis, M., Fila, G., Donatelli, M., 2002a. An indicator of solar radiation model performance based on a fuzzy expert system. Agronomy Journal 94, 1222-1233.

Bellocchi, G., Acutis, M., Paoletti, C., Confalonieri, R., Trevisiol, P., Grazioli, E., Delobel, C., Savini, C., Mazzara, M., van den Eede, G., 2008. Expanding horizons in the validation of GMO analytical methods: fuzzy-based expert systems. Food Analytical Methods 2, 126–135.

Bellocchi, G., Fila, G., Donatelli, M., 2002b. Integrated evaluation of cropping systems models by fuzzy-based procedure. 7th European Society for Agronomy Congress, 15-18 July, Cordoba, Spain, 241-242.

Bellocchi, G., Rivington, M., Donatelli, M., Matthews, K.B., 2010. Validation of biophysical models: issues and methodologies. A review. Agronomy for Sustainable Development 30, 109-130.

Bennett, N.D., Croke, B.F.W., Guariso, G., Guillaume, J.H.A., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T.H., Norton, J.P., Perrin, C., Pierce, S.A., Robson, B., Seppelt, R., Voinov, A.A., Fath, B.A., Andreassian, V., 2013. Characterising performance of environmental models. Environmental Modelling & Software 40, 1-20.

Bregaglio, S., Donatelli, M., Confalonieri, R., Acutis, M., Orlandini, S., 2010. An integrated evaluation of thirteen modelling solutions for the generation of hourly values of air relative humidity. Theoretical and Applied Climatology 102, 429-438.

Bregaglio, S., Donatelli, M., Confalonieri, R., Acutis, M., Orlandini, S., 2011. Multi metric evaluation of leaf wetness models for large-area application of plant disease models. Agricultural and Forest Meteorology 151, 1163-1172.

Carozzi, M., Bregaglio, S., Scaglia, B., Bernardoni, E., Acutis, M., Confalonieri, R., 2013. The development of a methodology using fuzzy logic to assess the performance of cropping systems based on a case study of maize in the Po Valley. Soil Use and Management 29, 576-585.

Confalonieri, R., Acutis, M., Bellocchi, G., Donatelli, M., 2009a. Multi-metric evaluation of the models WARM, CropSyst, and WOFOST for rice. Ecological Modelling 220, 1395-1410.

Confalonieri, R., Bellocchi, G., Boschetti, M., Acutis, M., 2009b. Evaluation of parameterization strategies for rice modelling. Spanish Journal of Agricultural Research 7, 680-686.

Confalonieri, R., Bregaglio, S., Acutis, M., 2010a. A proposal of an indicator for quantifying model robustness based on the relationship between variability of errors and of explored conditions. Ecological Modelling 221, 960-964.

Confalonieri, R., Bregaglio, S., Acutis, M., 2012. Quantifying plasticity in simulation models. Ecological Modelling 221, 159-166.

Confalonieri, R., Bregaglio, S., Bocchi, S., Acutis, M., 2010b. An integrated procedure to evaluate hydrological models. Hydrological Processes 24, 2762-2770.

Confalonieri, R., Gusberti, D., Acutis, M., 2006. Comparison of WOFOST, CropSyst and WARM for simulating rice growth (Japonica type – short cycle varieties). Italian Journal of Agrometeorology 3, 7-16.

Criscuolo, L., Donatelli, M., Bellocchi, G., Acutis, M., 2007. Component and software application for model output evaluation. Farming Systems Design 2007: an international

symposium on Methodologies for Integrated Analysis of Farm Production Systems, September 10-12, Catania, Italy, 2, 211-212.
Diodato, N., Bellocchi G., 2007a. Modelling reference evapotranspiration over complex terrains from minimum climatological data. Water Resources Research 43, doi:10.1029/2006WR005405.
Diodato, N., Bellocchi, G., 2007b. Modelling solar radiation over complex terrains using monthly climatological data. Agricultural and Forest Meteorology 144, 111-126.
Donatelli, M., Acutis, M., Bellocchi, G., Fila, G., 2004. New indices to quantify patterns of residuals produced by model estimates. Agronomy Journal 96, 631-645.
Donatelli, M., Rizzoli, A.E., 2008. A design for framework-independent model components of biophysical systems. In: Sànchez-Marrè, M., Béjar, J., Comas, J., Rizzoli, A., Guariso, G. (Eds.) iEMSs2008 International Congress on Environmental Modelling and Software, July 7-10, Barcelona, Spain, 2, 727-734.
Fila, G., Bellocchi, G., Acutis, M., Donatelli, M., 2003a. IRENE: a software to evaluate model performance. European Journal of Agronomy 18, 369-372.
Fila G., Bellocchi G., Donatelli M., Acutis M., 2003b. IRENE_DLL: a class library for evaluating numerical estimates. Agronomy Journal 95, 1330-1333.
Greenwood, D.J., Neeteson, J.J., Draycott, A., 1985. Response of potatoes to N fertilizer: dynamic model. Plant and Soil 85, 185–203.
Jain, S., Sudheer, K., 2008. Fitting of hydrologic models: a close look at the Nash–Sutcliffe index. Journal of Hydrologic Engineering 13, 981-986.
Laniak, G.F., Olchin, G., Goodall, J., Voinov, A., Hill, M., Glynn, P., Whelan, G., Geller, G., Quinn, N., Blind, M., Peckham, S., Reaney, S., Gaber, N., Kennedy, R., Hughes, A., 2013. Integrated environmental modelling: A vision and roadmap for the future. Environmental Modelling & Software 39, 3-23.
Legates, D.R., McCabe, G.J., 2013. A refined index of model performance: a rejoinder. International Journal of Climatology 33, 1053-1056.
Loague, K., Green, R.E., 1991. Statistical and graphical methods for evaluating solute transport models: overview and application. Journal of Contaminant Hydrology 7, 51-73.
Matthews, K.B., Rivington, M., Blackstock, K., McCrum, G., Buchan, K., Miller, D.G., 2011. Raising the bar? - The challenges of evaluating the outcomes of environmental modelling and software. Environmental Modelling and Software 26, 247-257.
Moriasi, D.N., Arnold, J.G., van Liew, M.W., Bingner, R.L., Harmel, R.D., Veith, T.L., 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Transactions of the ASABE 50, 885−900.
Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models, Part I - A discussion of principles. Journal of Hydrology 10, 282–290.
Olesen, H.R., Chang, J.C., 2010. Consolidating tools for model evaluation. International Journal of Environment and Pollution 40, 175-183.
Richter, K., Atzberger, C., Hank, T.B., Mauser, W., 2012. Derivation of biophysical variables from Earth Observation data: validation and statistical measures. Journal of Applied Remote Sensing 6, 063557.
Ritter, A., Muñoz-Carpena, R., 2013. Performance evaluation of hydrological models: Statistical significance for reducing subjectivity in goodness-of-fit assessments. Journal of Hydrology 480, 33-45.
Rivington, M., Bellocchi, G., Matthews, K.B., Buchan, K., 2005. Evaluation of three model estimations of solar radiation at 24 UK stations. Agricultural and Forest Meteorology 135, 228-243.
Stöckle, C.O., Kjelgaard, J., Bellocchi, G., 2004. Evaluation of estimated weather data for calculating Penman-Monteith reference crop evapotranspiration. Irrigation Science 1, 39–46.
Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. Journal of Geophysical Research 106, 7183-7192.

Tedeschi, L.O., 2006. Assessment of the adequacy of mathematical models. Agricultural Systems 89, 225-247.

Wallach, D., 2006. Evaluating crop models. In: Wallach, D., Makowski, D., Jones, J.W. (Eds.) Working with dynamic crop models: evaluation, analysis, parameterization, and applications. Elsevier, Amsterdam, 11-50.

Willmott, C.J., 1981. On the validation of models. Physical Geography 2, 184-194.

Willmott, C.J., Robeson, S.M., Matsuura, K., 2012. A refined index of model performance. International Journal of Climatology 32, 2088-2094.

Appendix 1

Multi-metric fuzzy-logic based evaluation of crop/grassland models in a model-intercomparison at multiple sites - Questionnaire

1) Do the fuzzy-logic based assessment method proposed (MQI_m), including model agreement, complexity and robustness, account for all the relevant aspects of multi-site model inter-comparison?

YES		**reason of "NO"**	**Proposal in case of "NO"**
NO			

2) Do the basic assessment metrics of MQI_m represent a good choice to cover aspects of model evaluation such as quantification of error, bias, efficiency, etc.?

YES		**reason of "NO"**	**Proposal in case of "NO"**
NO			

3) Do the equations of the basic metrics require changes (e.g. is standardized agro-meteorological metric, SAM, a good indicator of site conditions)? In case, how would you revise them to accommodate the needs of model evaluation, and why?

YES		**reason of "NO"**	**Proposal in case of "NO"**
NO			

4) Do the favourable/unfavourable trehsholds assigned to each basic metric reflect the perception of how we think of the quality of model performance?

YES		**reason of "NO"**	**Proposal in case of "NO"**
NO			

5) Do the expert weights assigned to basic metrics within a Module reflect the importance of each metric with respect to the quality of model performance?

YES		**reason of "NO"**	**Proposal in case of "NO"**
NO			

6) Do the expert weights assigned to Modules reflect the importance of each Module with respect to the quality of model performance?

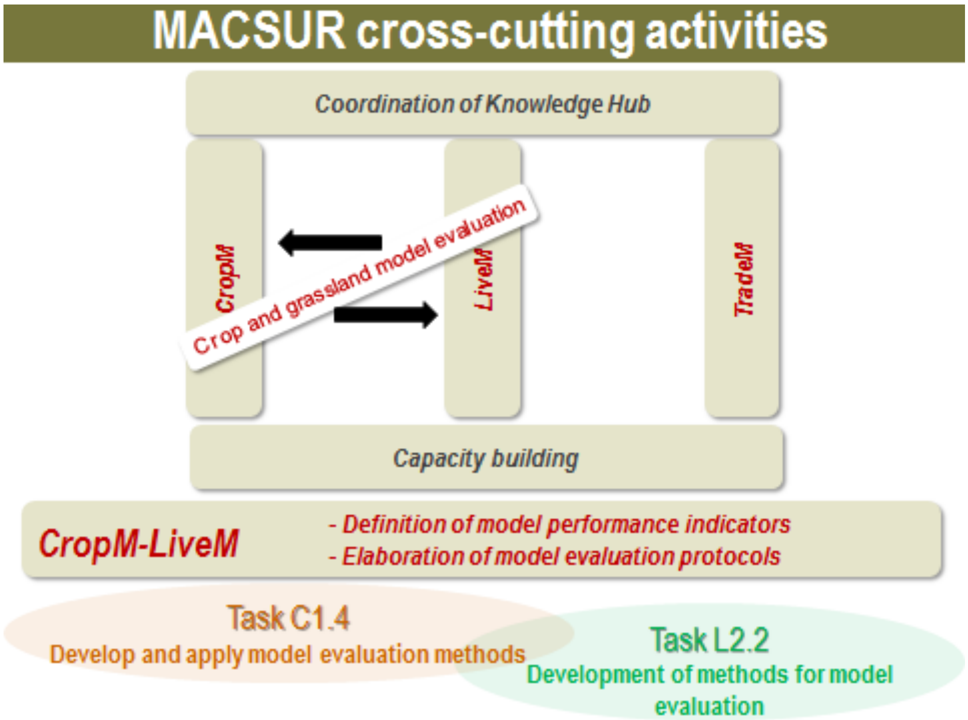
YES		**reason of "NO"**	**Proposal in case of "NO"**
NO			

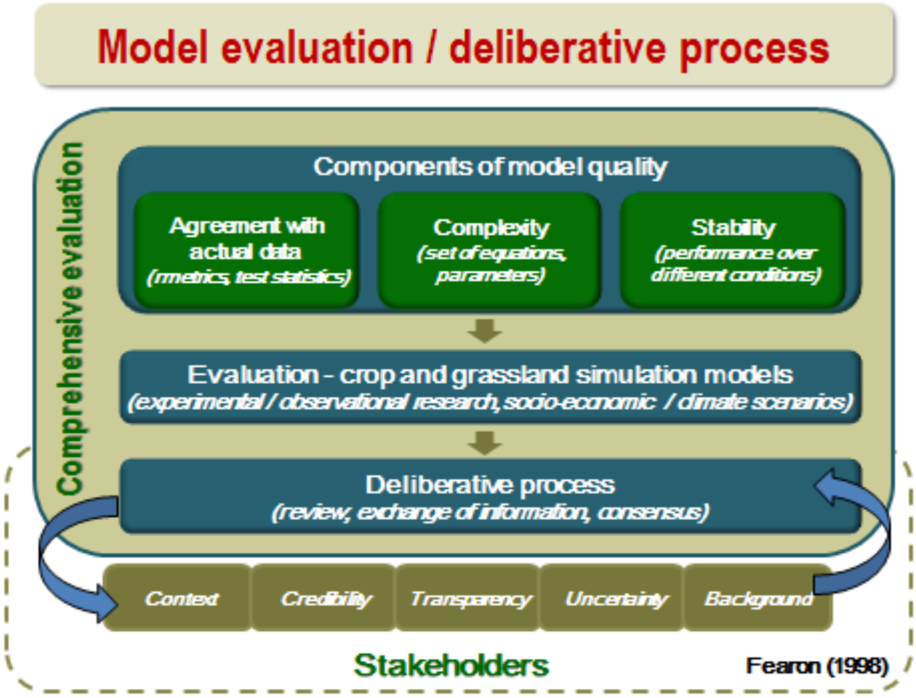
7) Over the range 0 (best) to 1 (worst) of MQI_m (and its three modules), would you set crisp threshold values to interpret results (e.g. <0.33: good model performance; 0.33-0.66: acceptable model performance but better calibration required; >0.66: poor model performance, improvements being required in the basic equation)?

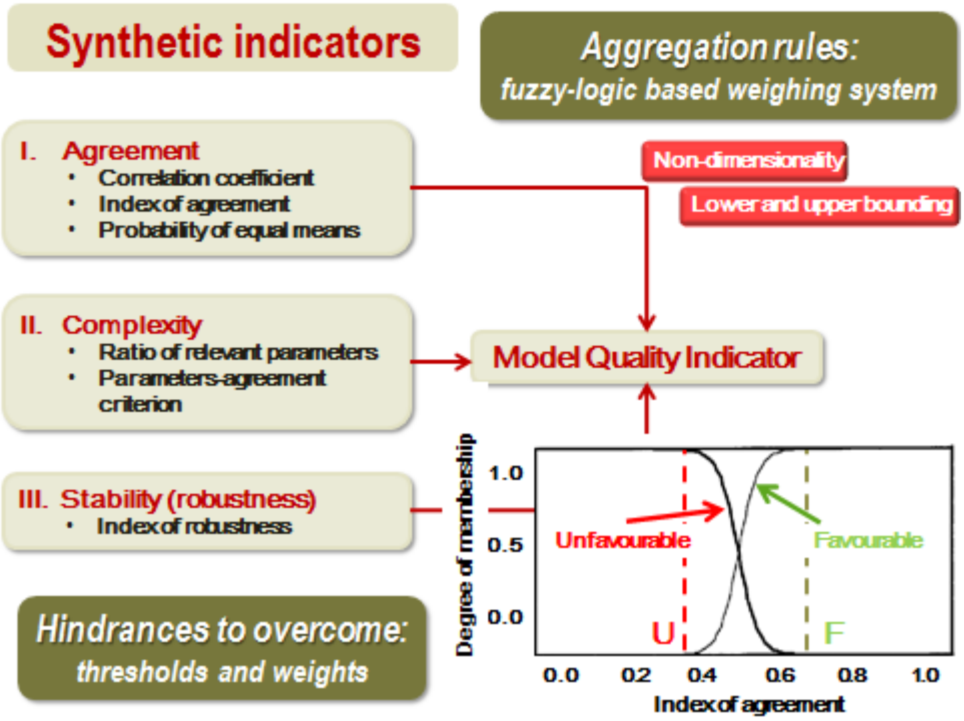
NO		**reason of "YES"**	**Proposal in case of "YES"**
YES			

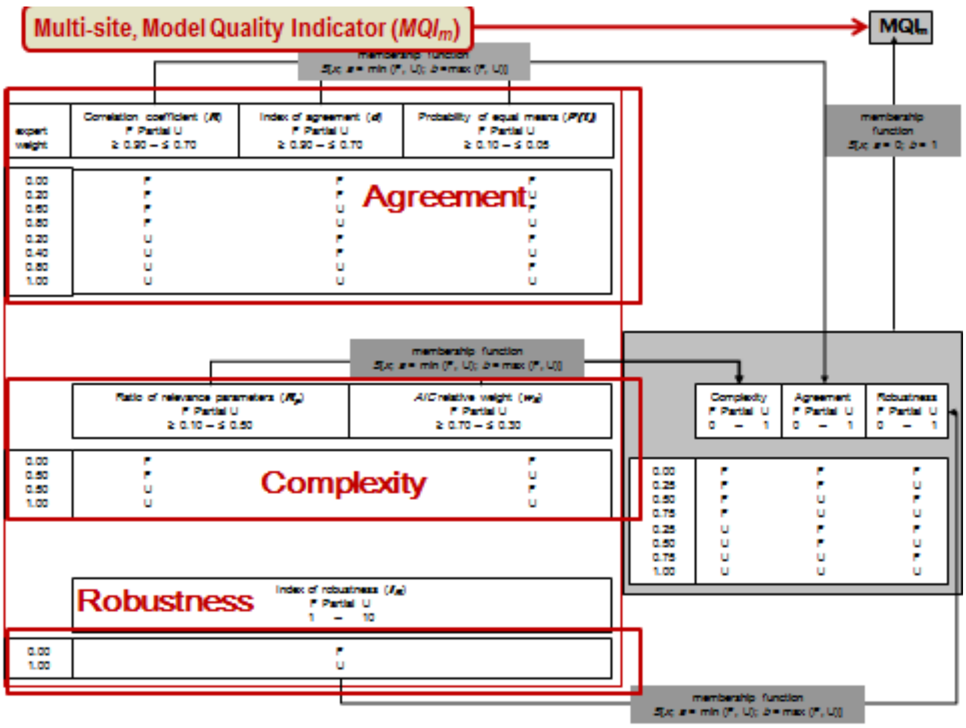
Appendix 2

Presentation given to FACCE MACSUR Mid-term Scientific Conference, 01-04 April, 2014, Sassari, Italy (http://ocs.macsur.eu/index.php/Hub/Mid-term/paper/view/193)


MACSUR cross-cutting activities
Coordination of Knowledge Hub
CropM
LiveM
TradeM
Crop and grassland model evaluation
Capacity building
CropM-LiveM
- Definition of model performance indicators
- Elaboration of model evaluation protocols
Task C1.4
Develop and apply model evaluation methods
Task L2.2
Development of methods for model evaluation


Model evaluation / deliberative process
Comprehensive evaluation
Components of model quality
Agreement with actual data (metrics, test statistics)
Complexity (set of equations, parameters)
Stability (performance over different conditions)
Evaluation - crop and grassland simulation models
(experimental / observational research, socio-economic / climate scenarios)
Deliberative process
(review, exchange of information, consensus)
Context
Credibility
Transparency
Uncertainty
Background
Stakeholders
Fearon (1998)

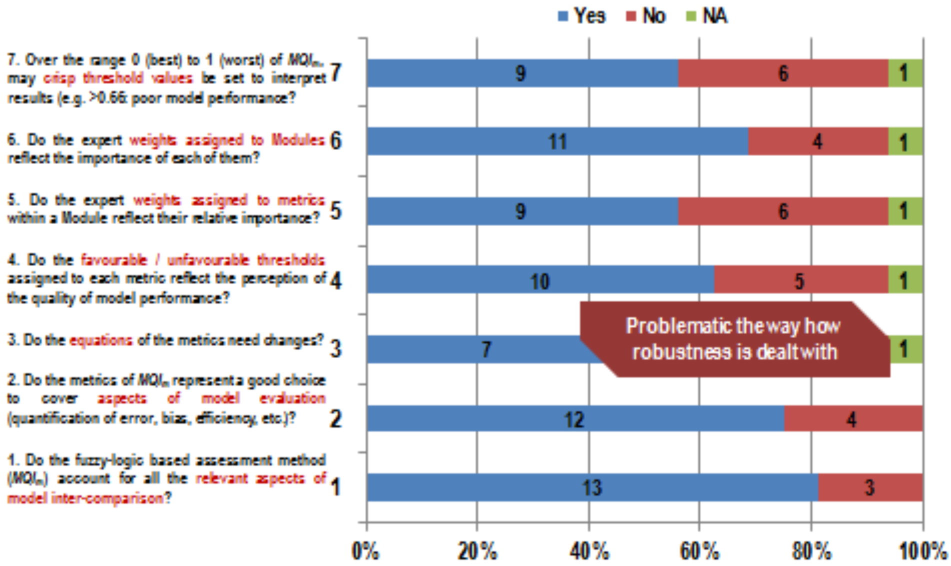

Synthetic indicators
Aggregation rules:
fuzzy-logic based weighing system
I. Agreement
• Correlation coefficient
• Index of agreement
• Probability of equal means
Non-dimensionality
Lower and upper bounding
II. Complexity
• Ratio of relevant parameters
• Parameters-agreement criterion
Model Quality Indicator
III. Stability (robustness)
• Index of robustness
Hindrances to overcome:
thresholds and weights
Degree of membership
1.0
0.5
0.0
Unfavourable
Favourable
U
F
0.0
0.2
0.4
0.6
0.8
1.0
Index of agreement



Multi-site, Model Quality Indicator (MQIm)
MQIm
Agreement
Complexity
Robustness

MQI_m – Questionnaire

Questionnaires answered / commented: 16 (13 online + 3 offline) + 1 comment



Robustness of a model

A **robustness measure** would account for model performance stability over a wide range of conditions (single site versus multiple sites)

How the variability of model performance can be quantified with the variability of conditions?

Index of robustness

Confalonieri et al. (2010)

$$I_R = \frac{\sigma_{EF}}{\sigma_{SAM}}$$ (0, best; +∞, worst)

Modelling efficiency

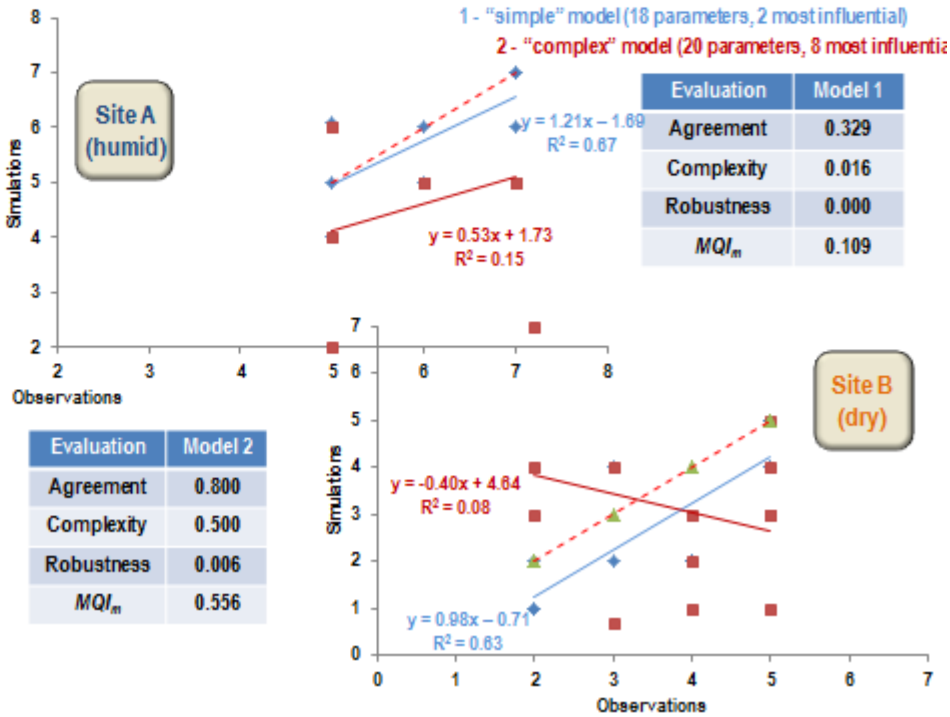
$$EF = 1 - \frac{\sum_{i=1}^{n}(P_i - O_i)^2}{\sum_{i=1}^{n}(O_i - \bar{O})^2}$$ (-∞, worst; 1, best)

Synthetic Agro-Meteorological Indicator

$$SAM = \frac{Rain - ET_0}{Rain + ET_0}$$ (-1, +1)

From the questionnaires:

- Need to test the index on a variety of rainfall patterns (e.g. monsoonal areas)
- Whole year versus growing season, or winter and summer?
- Accounting for soil properties if water limited simulations are performed



Evaluation	Model 1
Agreement	0.329
Complexity	0.016
Robustness	0.000
MQI$_m$	0.109

Evaluation	Model 2
Agreement	0.800
Complexity	0.500
Robustness	0.006
MQI$_m$	0.556

Exemplary results

Above-ground rice biomass (kg DM m^{-2})

Three models: WARM (intermediate), CropSyst (simple), WOFOST (complex)

MQI_s	WARM	CropSyst	WOFOST
C. d'Agogna	**0.0313**	0.1250	0.2174
Vercelli	0.1070	**0.0853**	0.1372
Mortara	0.2188	**0.0000**	0.2174
Rosate	**0.0313**	0.2284	0.2388
MQI_m	**0.0750**	0.1940	0.3356

EF	WARM	CropSyst	WOFOST
C. d'Agogna	0.90	0.95	0.93
Vercelli	0.92	0.97	0.96
Mortara	0.96	**0.98**	**0.98**
Rosate	0.92	**0.62**	**0.48**
I_R	**0. 16**	1.24	1.71

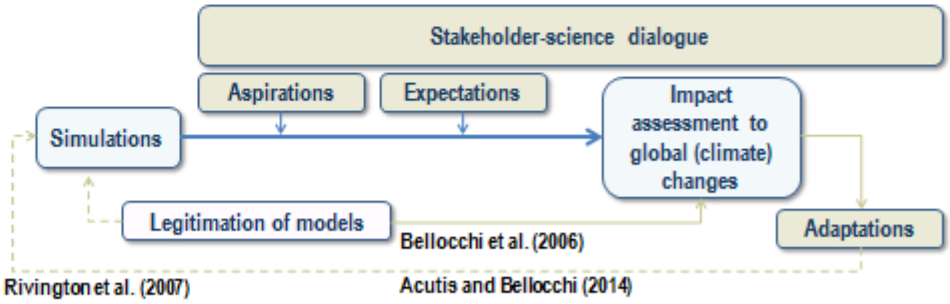
Robustness

MSE	WARM	CropSyst	WOFOST
C. d'Agogna	3.26	**1.86**	2.42
Vercelli	2.93	**1.35**	1.57
Mortara	1.66	**0.84**	0.94
Rosate	**0.97**	**4.96**	**6.75**

AIC	WARM	CropSyst	WOFOST
C. d'Agogna	**34**	37	**79**
Vercelli	**33**	34	**73**
Mortara	**26**	28	**67**
Rosate	**20**	49	**91**

Complexity

Deliberative process in model-based climate change studies



Implementation and resources / 1

... | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | ...

AgMIP

JPI FACCE : MACSUR, CN-MIP, ...

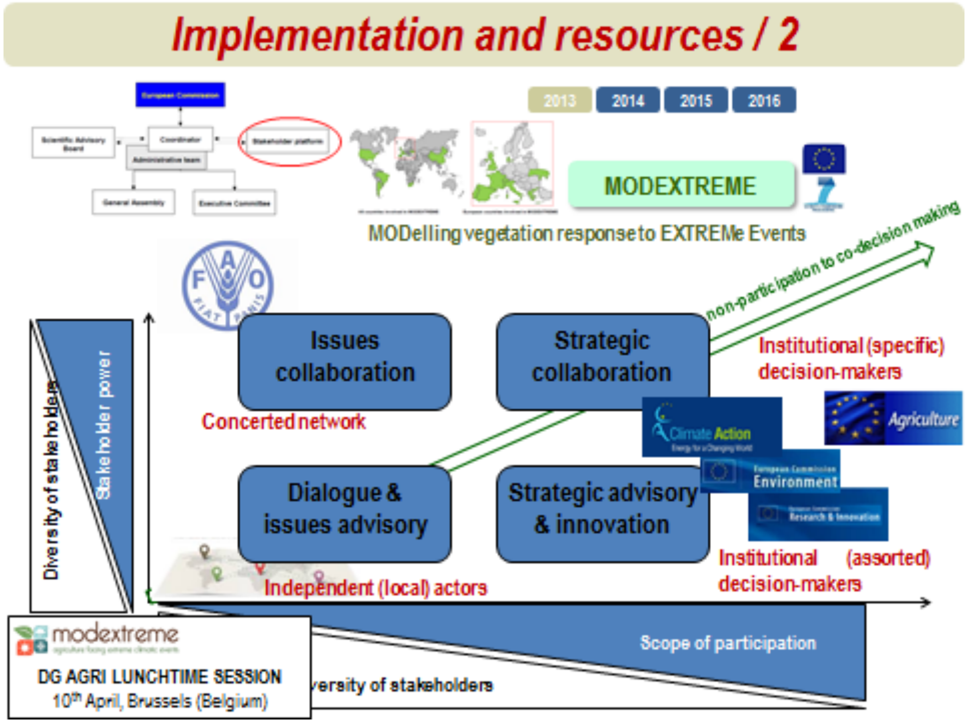
MACSUR knowledge hub (as well as parallel programmes such as AgMIP or other initiatives of the JPI FACCE) holds potential to advance in good modelling practice in relation with model evaluation (including access to appropriate software tools), an activity which is frequently neglected in the context of time-limited projects.

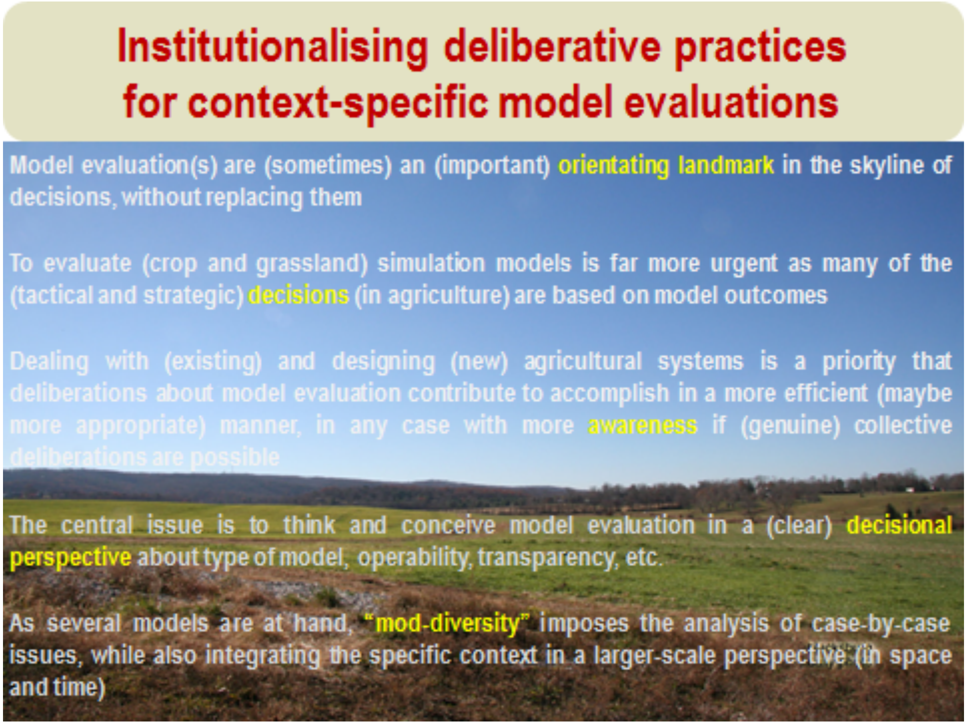
MACSUR

MACSUR Mid Term Conference
1st-4th April, Sassari (Italy)

LiveM →

International Livestock Modelling and Research Colloquium
14th-16th October, Bilbao (Spain)


Implementation and resources / 2
2013
2014
2015
2016
MODEXTREME
MODelling vegetation response to EXTREMe Events
non-participation to co-decision making
Issues collaboration
Strategic collaboration
Concerted network
Institutional (specific) decision-makers
Dialogue & issues advisory
Strategic advisory & innovation
Independent (local) actors
Institutional (assorted) decision-makers
Diversity of stakeholders
Stakeholder power
Scope of participation
Diversity of stakeholders
Climate Action
Agriculture
Environment
modextreme
DG AGRI LUNCHTIME SESSION
10th April, Brussels (Belgium)


Institutionalising deliberative practices
for context-specific model evaluations
Model evaluation(s) are (sometimes) an (important) orientating landmark in the skyline of decisions, without replacing them
To evaluate (crop and grassland) simulation models is far more urgent as many of the (tactical and strategic) decisions (in agriculture) are based on model outcomes
Dealing with (existing) and designing (new) agricultural systems is a priority that deliberations about model evaluation contribute to accomplish in a more efficient (maybe more appropriate) manner, in any case with more awareness if (genuine) collective deliberations are possible
The central issue is to think and conceive model evaluation in a (clear) decisional perspective about type of model, operability, transparency, etc.
As several models are at hand, "mod-diversity" imposes the analysis of case-by-case issues, while also integrating the specific context in a larger-scale perspective (in space and time)

"***We conserve many things that we don't evaluate and little of those we value***"
(Geoffrey M. Heal)

Appendix 3

Spreadsheet prototype for MQI_m calculation

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1			site 1			site 2			site 3			site 4			site 5	
2		obs	mod 1	mod 2	obs	mod 1	mod 2	obs	mod 1	mod 2	obs	mod 1	mod 2	obs	mod 1	mod 2
3		7,4	7,6	5,5	3,4	2,8	6,0	4,8	5,7	6,5	5,8	6,0	4,7	5,7	5,7	1,2
4		7,2	7,7	5,9	3,5	3,8	3,6	4,5	3,5	2,2	5,6	5,7	5,1	4,7	2,4	1,6
5		6,5	6,8	5,4	5,9	7,6	5,3	4,7	4,9	3,4	4,2	4,3	3,6	3,3	4,0	0,7
6		6,0	5,5	6,7	5,8	5,6	5,3	5,7	5,6	5,8	5,1	6,3	4,0	5,5	4,6	3,2
7		7,4	6,6	5,3	4,3	3,7	3,4	5,2	4,9	5,3	5,9	6,5	7,7	4,1	2,6	3,3
8		6,8	5,5	5,3	4,7	6,2	8,4	5,9	8,3	3,2	4,9	6,3	4,0	5,2	5,3	5,4
9		5,8	6,1	2,9	5,4	7,8	4,1	5,3	4,5	1,6	4,8	5,2	5,0	3,2	2,3	4,7
10		5,2	4,1	4,7	3,9	2,6	5,0	4,9	2,5	4,4	6,0	5,2	6,1	2,7	1,6	7,2
11					4,8	7,1	2,6	5,3	6,0	4,1	4,6	4,3	3,3	5,8	5,2	3,4
12					5,4	5,1	4,2	4,4	4,3	4,9	5,8	4,9	3,6	2,9	1,3	3,7
13					4,8	5,0	7,3	5,0	3,8	3,9	4,1	5,6	4,7	4,6	2,7	2,0
14					4,0	5,0	4,8				4,4	3,1	4,8	3,0	3,0	4,8
15					5,6	6,9	3,6				5,7	7,2	7,0	5,1	4,7	4,6
16					5,4	4,0	5,0				5,3	5,2	4,1			
17					3,4	3,2	4,5				4,5	5,8	3,9			
18					4,6	3,0	7,9									
19																
20																
21																
22	mean	6,5	6,2	5,2	4,7	5,0	5,1	5,1	4,9	4,1	5,1	5,4	4,8	4,3	3,5	3,5
23																
24																
25	**Pearson correlation**		0,820	0,394		0,864	0,101		0,688	0,067		0,512	0,724		0,769	-0,389
26	**coeff of agreement**		0,846	0,488		0,743	0,275		0,554	0,260		0,634	0,635		0,800	0,221
27	**Student T (P for paired data)**		0,262	0,011		0,388	0,407		0,663	0,071		0,179	0,284		0,007	0,287

Data | Thresholds | Modules | Modules aggregation | DataGenerator

Prêt 115 %

	A	B	C	D	E	F	G	H	I	J	K	L
1												
2	**METRICS**											
3			THRESHOLDS		Weight within a module		Average of all sites		F values		U Values	
4							M1	M2	M1	M2	M1	M2
5			U	F								
6	Module Agreement											
7		Pearson correlation coefficient	0,7	0,9	0,2		**0,731**	**0,179**	0,047	0,000	0,953	1,000
8		coefficient of agrreement d	0,7	0,9	0,6		**0,716**	**0,376**	0,012	0,000	0,988	1,000
9		Student-t for paired data (P(t))	0,05	0,1	0,2		**0,300**	**0,212**	1,000	1,000	0,000	0,000
10												
11	Module Complexity											
12		relevant over total parameters ratio (Rp)	0,5	0,1	0,5		**0,125**	**0,667**	0,992	0,000	0,008	1,000
13		weighed measure (wk) of the AIC	0,3	0,7	0,5		**0,980**	**0,020**	1,000	0,000	0,000	1,000
14												
15	Module Robustness											
16		index of robustness (IR)	10	1			**13,65**	**25,76**	0,000	0,000	1,000	1,000
22		user input										
23		Could be calculated in the sheet "data" or user input										

$S(x;\alpha;\gamma)$

Data | Thresholds | Modules | Modules aggregation | DataGenerator

Prêt 100 %

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1									Membership value								
2	**Module Agreement**								**M1**					**M2**			
3	Expert weight	Pearson correlation coefficient	coefficient of agrreement d	Student-t for paired data (P(t))	to compute expert weight				Pearson	d	Student t	Thruth val	Thrut*wei	Pearson	d	Student t	Thruth v
4	0,0	F	F	F	0	0	0		0,047	0,012	1,00	0,012	0,000	0,00	0	1	0,00
5	0,2	F	F	U	0	0	0,2		0,047	0,012	0,00	0,000	0,000	0,00	0	0	0,00
6	0,6	F	U	F	0	0,6	0		0,047	0,988	1,00	0,047	0,028	0,00	1	1	0,00
7	0,8	F	U	U	0	0,6	0,2		0,047	0,988	0,00	0,000	0,000	0,00	1	0	0,00
8	0,2	U	F	F	0,2	0	0		0,953	0,012	1,00	0,012	0,002	1	0	1	0,00
9	0,4	U	F	U	0,2	0	0,2		0,953	0,012	0,00	0,000	0,000	1	0	0	0,00
10	0,8	U	U	F	0,2	0,6	0		0,953	0,988	1,00	0,953	0,763	1	1	1	1,00
11	1,0	U	U	U	0,2	0,6	0,2		0,953	0,988	0,00	0,000	0,000	1	1	0	0,00
12																	
13												1,024	0,793				1,00
14	Module weight =	0,5															
15												I=	**0,774**				I=
16																	
17																	
18	**Module Complexity**								**M1**					**M2**			
19	Expert weight	relevant over total parms ratio (Rp	weighed measure (wk) of the AIC						Rp	WkAIC		Thruth val	Thrut*wei	Rp	WkAIC		Thruth v
20	0,0	F	F		0	0			0,992	1,000		0,992	0,000	0	0		0,00
21	0,5	F	U		0	0,5			0,992	0,008		0,008	0,004	0	1		0,00
22	0,5	U	F		0,5	0			0,008	1,000		0,008	0,004	1	0		0,00
23	1,0	U	U		0,5	0,5			0,008	0,008		0,008	0,008	1	1		1,00
24																	
25												1,016	0,016				1,00
26	Module weight =	0,25															
27												**I=**	**0,015**				**I=**
28																	
29																	
30	**Module Robustness**																
31	Expert weight	index of robustness (IR)															
32	0	F															

Data | Thresholds | Modules | Modules aggregation | DataGenerator

Prêt 100 %

	Modules							M1					M2				
Expert weight	Agreement	Complexity	Robustness	To compute weights				Agreement	Complexity	Robustness	Thruth value	Thrut*weight	Agreement	Complexity	Robustness	Thruth value	Thrut*weight
0	F	F	F	0	0	0		0,774	0,015	1	0,015	0,000	0,800	0,500	1	0,500	0,000
0,25	F	F	U	0	0	0,25		0,774	0,015	0	0,000	0,000	0,800	0,500	0	0,000	0,000
0,25	F	U	F	0	0,25	0		0,774	0,985	1	0,774	0,194	0,800	0,500	1	0,500	0,125
0,5	F	U	U	0	0,25	0,25		0,774	0,985	0	0,000	0,000	0,800	0,500	0	0,000	0,000
0,5	U	F	F	0,5	0	0		0,226	0,015	1	0,015	0,008	0,200	0,500	1	0,200	0,100
0,75	U	F	U	0,5	0	0,25		0,226	0,015	0	0,000	0,000	0,200	0,500	0	0,000	0,000
0,75	U	U	F	0,5	0,25	0		0,226	0,985	1	0,226	0,169	0,200	0,500	1	0,200	0,150
1	U	U	U	0,5	0,25	0,25		0,226	0,985	0	0,000	0,000	0,200	0,500	0	0,000	0,000
											1,031	0,370				1,400	0,375
											MQI=	**0,359**				MQI=	**0,268**

Data / Thresholds / Modules / **Modules aggregation** / DataGenerator

site 1		site 2		site 3		site 4		site 5	
obs	mod	obs	mod	obs	mod	obs	mod	obs	mod
5,68802	8,28198	3,06601	2,86411	5,54177	3,76871	5,79653	3,93302	5,3551	7,28753
7,79729	3,93704	3,07074	4,50895	5,52767	7,0867	4,7882	4,94483	2,26028	0,8614
7,99689	9,10478	3,8233	4,08226	4,43267	3,50157	5,53578	3,3339	5,53707	7,16308
7,52462	7,202	4,27523	5,70795	5,06043	3,09024	5,50739	4,83925	3,48616	4,23513
7,69212	5,80267	4,71933	3,58796	5,30631	6,35199	4,94569	3,59547	5,94868	3,02041
6,5348	5,96204	5,99974	6,73007	4,99779	2,95428	5,71587	8,47169	4,61877	5,64535
6,10437	7,82645	5,50808	4,24069	4,70596	6,42689	4,29759	6,34031	4,22385	2,88111
7,12556	4,41411	4,62972	7,23473	4,53469	6,61244	5,41189	7,79149	2,97914	1,92234
		5,45676	5,20931	4,04323	4,93258	4,46406	6,01101	5,37126	4,34619
		3,599	4,11954	4,47322	4,33741	4,35276	4,18063	4,6583	4,11119
		3,78842	3,67262	5,64355	6,50082	5,86739	5,86271	5,89556	8,38595
		3,54342	2,78779			5,39447	6,42856	2,29336	1,41921
		4,08383	3,8422			5,95858	4,06374	2,11997	2,0401
		3,62599	4,46202			5,11849	3,95307		
		4,93415	6,03611			4,30535	3,27026		
		3,77244	6,00132						

Data / Thresholds / Modules / Modules aggregation / **DataGenerator**