Hydrol. Earth Syst. Sci. Discuss., 6, C3230–C3245, 2010
www.hydrol-earth-syst-sci-discuss.net/6/C3230/2010/

Hydrology and Earth System Sciences Discussions

Interactive comment on "Towards automatic calibration of 2-dimensional flood propagation models" *by* P. Fabio et al.

P. Fabio et al.

fabio@idra.unipa.it

Received and published: 29 January 2010

The authors wish to thank M. Horritt, M. Wilson and J. Neal for the fruitful review and the useful suggestions. Every comment gives us the opportunity to improve the manuscript. We will address each point below. For easier comprehension, each referee's comment is reported in italics.

Reply to M. Horritt

- In equation (4), Q represents the inverse covariance matrix - this would imply that the method takes into account both variable errors in the observations, and the correlation between them. A major difficulty in calibration studies is dealing with correlations in the

observed data - this tends to place more weight on areas where more measurements have been acquired, which may not be justified since the errors in these measurements may be strongly correlated. This method appears to offer a way to deal with this - have the authors looked at correlation between the errors, or assumed them to be independent?

The authors agree that it is very difficult to deal with correlations in observed data, especially when there is no a priori information about them. Also, the notion that this method might offer a solution for the problem of correlation in measurement errors is indeed an interesting future perspective. However, at present the measurements errors have to be assumed to be independent because the currently implemented algorithm in PEST assumes this explicitly (Doherty, 2004). This is expressed in equ. (4) by imposing strict diagonality of matrix Q. To our knowledge, this assumption is inherent in all automatic calibration routines.

- The authors calibrate the model against water depths rather than water elevations - this will have a large effect on the sources of error. Hydraulic models are generally much better at predicting water levels than at predicting water depths - the latter are much more sensitive to topographic errors. This may explain the spatial distribution of errors, which coincide with steeper slopes on the floodplain (these are the places where locational or sampling errors will have greatest effect). How would the results be affected by using water levels rather than depths?

We used water levels because we estimate the errors of the direct surveys of the maximum inundation depths as comparatively low as compared to the errors in the underlying digital elevation model. (A detailed discussion on data quality and errors is lead in the reply to Matt Wilsons review.) The survey errors can be estimated by the accuracy of the GPS survey point locationing and the measurement of the inundation depth above ground by laser distance meter. These can be quantified at approx. 6 m in location and below 1 cm for inundation depth. The errors in the 25 m resolution DEM are estimated at below 1m in the flood plains and up to several meters on

steep hillslopes. These errors would have been introduced in the measurement data by calibrating against water elevations. This is the main reason for using inundation depths instead of elevations for calibration. Basically we argue that ideally the inundation model should predict the inundation depths and not compensate for errors in the model setup and DEM. Additionally, from the view of the optimisation routine, it doesn't make a difference whether inundation depths or elevations are used, given that the observed inundation elevations are based on the sum of surveyed inundation depths and the ground elevation given by the DEM: the objective function, i.e. the sum of squares (equ. 5 6) evaluates to the same value: ïĄěi = hobs – hsim = (hobs + DEM) – (hsim + DEM). Thus we conclude that the parameter estimation would not be affected by using either observed inundation depths or inundation elevations. A possible work-around would have to survey the elevation of the surveyed points directly by high precision differential GPS. With this information the identification of DEM errors would have been possible a priori, as well as the use of inundation elevations in the calibration. This should be the preferred way of inundation surveys, especially when no high-precision LiDAR DEM is available.

- What is the length of the model reach? Fig 3 indicates it's 6-7km. My concern is that the reach is short, and therefore the dowsntream boundary condition (the authors should describe this) will have a significant influence on water levels throughout the reach (see Horritt et al, 2007, Comparing the performance of a 2-D finite element and a 2-D finite volume model of floodplain inundation using airborne SAR imagery, Hydrological Processes, 21(20), 2745-2759, for a discussion of this for a similar type of reach). If the downstream boundary condition is wrong, it may need to be compensated for by unrealistic roughness parameters. How can the authors be sure this is not happening in this study?

The length of the model reach is about 8.5 km. As described in section 2.2 the model runtime is about 4 hours, to consider a longer river length would be very costly in terms of model setup and model runtime. Moreover, all the calibrations performed would be

too much computationally demanding and could not be achieved. For this hydraulic model, boundary conditions are always given by the incoming unit flux along the upper part of the boundary and the water surface elevation along the lower part of the same boundary (Aronica et al., 1998b). As described in section 3, the data recorded by the downstream gauging station could not be used, thus, for the historical event, records or informations about downstream boundary conditions are not available. Therefore we tested different downstream boundary conditions and observed that there is indeed an influence on the downstream part of the river domain on the results, as the reviewer indicated. As lower boundary condition, in the model version used in the calibration we assumed normal water depth for the channel and zero water depths for the downstream nodes located in the floodplain. Reviewing the estimated roughnesses, which are comparatively high, it is quite likely that they compensate in part the model errors introduced by these simplifying lower boundary conditions. However we believe that these effects are justifiable, because the estimated parameters are effective parameters by any means and the focus of the study was on the testing of automatic calibration routines for hydrodynamic models, not on finding "real" roughness parameterisations.

- The authors are right to be cautious about rejecting observations because they don't fit the model (it's more likely that the model is wrong). Effectively, observations are rejected if they are not included in the range of model predictions from the range of input parameters. This type of behaviour is generally associated with an inadequate model - is there any evidence for large model errors that may cause this? I notice there are two bridges across the channel (I think!) - are these represented in the model? How confident are we in the hydrometric data? The authors should include a discussion of these possible error sources.

As discussed in the introduction of the paper, the authors are certain that a computer-based model is an imperfect representation of a physical system and that a perfect match is not expected from a calibration to the available field measurements, due to the presence of errors both in data and in the model. However, we assumed that the

mathematical structure of the model, i.e. the mathematical model is predetermined and fixed and that also hydrometric data are correct, which can admittedly by discussed, especially for such a large event. However, we put some confidence in the data because they were error checked and corrected by the hydrometric authorities of the region operation the gauges and providing the data. The large upstream railway bridge is represented in the model, whereas the smaller downstream bridge is not. This most likely introduces some errors, but we assume these small compared to the errors introduced by the DEM. In fact, we define the DEM as part of the hydrodynamic model. Thus our proposed method of identifying errors in the model follows exactly the notion of the reviewer, that the observations cannot be explained by the predictions of the model.

- Some typos etc to fix: P6833: Line 11: Suggest change to "Secondly, two dimensional models in particular..." Line 16: The model can't be described as a full 2D model since it neglects the advection terms - suggest calling it a simplified 2D model instead. P6834: Line 16: "efficient" should be "effective" Line 17: Suggest change "praxis" to "practice" Line 25: Optimisation methods such as these have previously mostly been applied to hydrological models where parameters can be less well defined (ie less physically based) - the authors should mention this. P6835: Line 5: Suggest change "non-linear" to "complex". P6836: Line 1: Suggest change "hardly ever" to "rarely do" P6837: Line 11: Change "explicit" to "varying" P6838: Line 6: This sentence makes little sense - rewrite. P6846: Line 19: I'm not sure what the authors mean by "chroncially" - "typically"? P6848: Line 21: Change to "orthogonally".

The authors will include the suggested changes. As regards *"P6846: Line 19: I'm not sure what the authors mean by "chroncially" - "typically"?"* Yes, this is the meaning: errors in the DEM in this resolution are generally high on slopes and even more on the transition of flat areas to slopes. We say chronically, because we want to express typically, but with a negative touch.

Reply to M. Wilson

The validation data which are available are point based measurements of maximum flood depth and imagery of flood inundation extent. Because the flood event was a large valley-filling event, and the valley studied was topographically well contained, the flood extent data were not used since changes in predicted flood depth do not make any substantial difference to the area of inundation. This is appropriate, although it would still be good to show these data in a figure and illustrate how well the model predicts inundation extent. It is a shame that it wasn't a smaller event which would have allowed these data to be included as a more rigorous test. Without them, the study is limited to using the point measurements of maximum flood depth, which would be ok if they were spread across the whole domain, but unfortunately they only based in the urban area. Therefore, it is entirely unsurprising that the spatially distributed friction values from outside of this area only have a limited effect on accuracy and hence display equifinality. It may also be better to calibrate against water elevations rather than water depth – this would avoid minor errors from the DEM creeping into the calibration. Can the measured depths be converted to water elevation?

Matt Wilson is true in saying that it is a shame that it wasn't a smaller event (people from Eilenburg would also agree!), so that the inundation extent could have been used additionally. The initial intention was to use both inundation extend data and surveyed inundation depths in the automatic calibration. But because the inappropriateness of the inundation extend was already shown in a previous study using different hydraulic models and manual calibration on the event, we excluded the inundation extend in this study. The model results of estimating the inundation extend were published in Apel et al. (2009). The three models of different complexity (water level interpolation and DEM intersection, Lisflood-FP and the model presented here) all reached a Flood Area Index of 96Regarding the use of water elevation, please see the above answer to M. Horritt. In short, it can be done by adding the DEM elevation to the surveyed depths, but this would not change the results. For additional gain of the results an independent high precision altimetry by differential GPS would have been necessary, but that was unfortunately not performed. For the discussion of distributed roughness,

see discussion below. Apel, H., Aronica, G., Kreibich, H. and Thieken, A. (2009) Flood risk analysesâĂŤhow detailed do we need to be? Natural Hazards, 49(1): 79-98

Currently the paper is trying to do too much with a limited dataset. To this end, some simplifications are needed in the paper. I suggest the following either (i) only look and channel/ floodplain friction, therefore avoiding spatially distributed friction – this would reduce the equifinality observed but still illustrate the principles of automatic calibration, which is the main focus of the paper; or (ii) reduce the domain size so that it focuses on the urban area and the calibration points are then spread more completely across the domain – this would again reduce the equifinality; however, the domain size may already be rather small and care is needed to avoid boundary effects. Or if additional data for a second event were available, they would be very helpful – although this seems rather unlikely.

First of all we disagree to the statement that the data set used is limited. It is in fact very extensive - a study with a similar number of surveyed inundation depths has not been published before. The dataset itself is indeed limited to the urban area, but this does not restrict the use of distributed roughnesses. We have limitations in the accuracy of the DEM, but with the proposed method we could identify them. We elaborate these pints in the following starting with description of the data used and their estimated errors: DEM: The 25m resolution DEM used in the study was generated and issued from the German Federal Authority of Geodesy and Cartography and is based on topographical maps with a scale of 1:25000. These maps, especially the elevation information are based on terrestrial surveys. The accuracy of the elevation of this product is given as +/- 1m in lowland areas and up to +/- 7m in mountainous regions. The latter estimation has to be assumed for the steep hillslopes bounding the valley floor in the study area. Inundation depths: The inundation depths were surveyed by laser reflectometry of water marks on buildings above ground. The accuracy of the laser measurements itself is in the mm-order. Some higher errors are introduced when the ground reference level was hard to determine, e.g. by bushes growing in front of the building or unclear water marks.

No detailed information is available about this, but the errors of the surveyed maximum inundation depths can be assumed at below 0.2m. The location of the survey points was determined by handheld GPS. Thus the errors in the location of the survey points are below 6m. Comparing these accuracies it becomes clear what we used inundation depths rather than inundation levels for the calibration. Land use: We used the official CORINE land use classification, which is a European standard. Thus the selection of the land use was not arbitrary, but based on the standard classification. The full land use classification is equivalent to the 5 parameter calibration, where floodplain is equivalent to the original classification as meadow/grassland. All calibrations with lower number of parameters are aggregations of these classes. The calibration scenarios B and C correspond to the aggregation suggested by Matt Wilson. The scenario D with 4 parameters is also not arbitrary but based on the assumption that flow through a mostly paved urban environment, especially with high flow depths as in this case, is similar to open channel flow, thus we aggregated channel and urban area. We are aware that this assumption is debatable and other, in fact opposite approaches are used and published. Therefore this aggregation level could be dropped. However, we argue that using the full distributed land use is of value for this study, because of the following reasons: (1) the calibration is not dominated by the roughness of the urban area, but also of the surrounding floodplain. This is evident in Table 2 through the 95(2) Next to the main point of this paper, the study on the feasibility of automatic calibration routines in hydraulic modelling, we want to show that equifinality can also be caused by an inverse mismatch in data and model complexity: a too simple model setup compared to the number of available calibration data. By removing the different calibration strategies we would not reduce the equifinality, we would simply not illustrate them. With the proposed calibration layout we could identify 2 model setups that perform equally well (B F) and discard others. This we regard as valuable information of the paper.

Specific comments: P6836, L17-22. The selection of these roughness scenarios seems quite arbitrary. It may be better to use a true continuous friction value based on the land cover. See for example Mason et al. (2003, Hydrological Processes 17(9),

1711–1732) or Wilson and Atkinson (2007, Hydrological Processes 21 (26) 3576-3596) for possible methods. These continuous surfaces would provide for a more rigorous approach in defining spatially distributed friction.

As discussed above, the selection of the roughness scenarios is not arbitrary, but based on the CORINE land use classification. Each land use class gets is assigned with an individual roughness value. The CORINE land use classes were further aggregated on three levels, from which all but the scenario D are standard roughness parameterization approaches. Thus it follows the same approach as in the references cited, just on a larger scale.

P6838, L14. What is the source of the DEM used? LiDAR? You provide a scale of 1:10,000, but scale is only one aspect – precision of the data are very important. 1:10,000 contours are still not good if only every 5 m. Please clarify what topographic data you are using, and provide an estimate of their likely accuracy.

Please see reply on general comment above.

P6841. Comment on the possible use of computational clusters for PEST analysis.

Computational clusters can well be used with PEST, because it provides a parallelization option. In fact we used the parallel version of PEST in combination with an 8-processor compute server. Also office grid solution would be possible. However, at present PEST is restricted to Windows-based operation systems, thus the usual Linux-based large computational clusters cannot be used unless Windows emulating software is installed.

P6842, L21-25. Similar to with the roughness scenarios, the dividing of the land use into principal regions is open to subjectivity. What process did you use to achieve it? This needs greater discussion since it is a fundamental part of the study.

Again, the CORINE land use classification was used. For detailed informations on methodology etc. refer to http://www.eea.europa.eu/publications/COR0-landcover

P6844, L17. Why didn't you exclude the unphysically realistic negative parameter values from the calibration set?

Those values were not actually used in the calibration model runs. The confidence bounds are derived from the distribution of the parameters used during the calibration. In case of low mean and high variance, the 95

P6844, L18-21. This discussion regarding the parameter values and equifinality would be better if put into the context of accuracy – so probably better moved to the discussion.

We will follow this suggestion.

P6844, L25-29. Expand on this brief discussion of the use of PAR2PAR – this appears to limit you somewhat.

PAR2PAR can be used to establish relationships between parameters, like the relationship adopted by us that the roughness in the floodplain, i.e. all land use classes outside the channel, should always be higher than in the channel. This is done by defining e.g. for the floodplain a "new" Strickler roughness parameter given by the ratio between the Strickler roughness coefficient of the channel and the Strickler roughness coefficient roughness of the floodplain, and imposing as lower bound of this parameter the unit value. The manual (Doherty, 2004) furnishes a very similar example referring to parameters governing infiltration of water into different parts of a catchment. However, the way PAR2PAR is implemented brings some restrictions on the calibration efficiency, as the results show. Moreover, it limited us in the definition of parameter bounds, because we can only define bounds for the "new" parameters. Therefore the further calibrations (with removed data) were performed without conditioning parameter.

P6845, L2-5. The other calibration criteria should be illustrated here.

We will elaborate on discharge hydrographs, maximum inundation extend, and spatially distributed water depths in the introduction and refer her accordingly.

P6845, L19-21. The RMSE is quite high (0.8 m) and very similar for all. What is the error in the measurements used? You could illustrate the calibration behaviour using dotty plots.

For a discussion on the errors of the measurements see our reply to the general comments. Given the errors associated to the DEM, the RMSE of 0.8m is already in the range of those errors. As shown in the example dotty plot below, the high RMSE is dominated by the outliers, where the simulated water depths drastically overestimate the surveyed. As argued in the discussion, this overprediction cannot be explained or rectified by roughness parameterization, this is rather a model error, i.e. the DEM. We will include some dotty plots in the revised manuscript to illustrate this. The reduced RMSE of the calibration using a reduced number of calibration point corresponds to a reduced overall error of the DEM at the surveyed locations.

P6845, L24. The roughness values are also suppressed because only maximum flood depth is used as comparison – this misses much of the dynamics which friction values will have an influence on.

This is correct. We will add a notion on this.

P6847 – It is difficult to assess quality of DEM here since we don't know where it came from or how it was produced. We also don't know how the calibration points were obtained and their likely error, so we don't have any idea of the relative error in each dataset. These should both be discussed earlier in the paper. But in any case, I don't think that removing calibration points is appropriate here, and doesn't somehow balance the model complexity with available data. It is always better to have as many points in as possible, which then may highlight the shortcomings of the model, as they have in this case. I suggest skipping this last step as I don't think it adds anything to the paper.

As laid out above, we don't remove calibration points because we think they are wrong. We remove them because they prevent better model calibration because of the errors

in the DEM at these points. And because we cannot fix these errors on an objective basis and the fact that these errors dominate the objective function, thus stalling the optimisation routine, we strive to obtain a better calibration for the remaining points by excluding them. Instead of removing them we could have assigned them with low weights, but the effect would be the same. The proposed method also serves for the identification of the most sensitive calibration points as illustrated in Table 4, which is also a valuable information. Also, skipping this part of the manuscript would mean skipping the part that illustrates how erroneous model setup influence the quality of the calibration and how those points can be identified on an objective basis (coefficient of variance over different roughness parameterization). We hope that these explanations convince the reviewer and the editor and given the fact that the other reviewers don't criticise the approach, would rather keep this part in the manuscript.

Minor points/ typos, etc:

The authors will include the suggested changes.

Reply to J. Neal

The paper presents the application of a gradient based automated calibration routine to a 2D hydraulic model. It then discusses the influence of errors associated with using depth observations and a 25 m resolution DEM on the calibration procedure. These errors are believed to be a significant cause of parameter equifinality and to adversely effect model calibration. If possible, a comment on the implications of these findings for future data collection initiatives that might then use automated calibration routines would be interesting. Specifically, the use of depth observations seems to have been problematic here?

Jeff Neal is right in saying that errors in both observation data and DEM can influence the calibration process by dominating the objective function and thus "stalling" the optimisation. In our case the dominating errors were with the DEM, cf. the discussion on data quality in the reply to Matt Horritt comments. The implications for automatic

calibration are that much care has to be taken in collecting data and, maybe even more possible, that the data collection is recorded properly in order to identify and quantify possible error sources. As laid out in the reply to Matt Horritt comments, the use of inundation depths were not the problematic. In fact, we argue that using inundation elevations for calibration is only useful if the ground elevation can be determined with high accuracy, e.g. by differential GPS ground surveys or LiDAR derived DEMs.

Would it be worth plotting CV against some simple metrics such as local DEM slope, distance from the channel etc. and could this information form a physical basis for rejecting/keeping observations? The paper suggests it wouldn't but a plot would add detail. As depth and a 25 m DEM are being supplied as observations it seems likely that these values will be of poor quality in steeply sloping areas (as pointed out in the text).

We agree and will investigate if a plot DEM slope against CV will illustrate this more clearly than the spatial arrangement.

In this sense does depth information present a similar problem to extent information in that its not very useful in steep areas?

No, as long as you've got the topography right and in an appropriate resolution. The problem here is not only the errors associated with the DEM, but also the resolution of the DEM. With this resolution the transition from the more or less flat floodplain to the steep hillslope cannot be represented properly. Because of the resolution the onset of the hill slopes is not gradually but discrete, i.e. like a wall some meters high.

Were there any problems when using the automated calibration that might require expert knowledge or additional simulations? In the introduction the gradient based method was criticised for having the potential to find local minima rather than global optimal parameter sets. Was this a problem here, did this change with the number of parameters in a set?

No further expert knowledge is required in using PEST or automatic calibration in general, but an understanding of the principles of optimisation is certainly helpful. The manual of PEST gives adequate advice on both. The problem of finding local minima was also investigated after this study and is subject of another planned publication. In short, we found that the gradient based method was able to find the global minimum.

Specific comments:

The authors will include the suggested changes. Below some explanations can be found.

P6834 L5-10: There has been quite a lot of research on the factors which introduce errors into inundation models, including roughness. With this in mind is it worth very briefly mentioning some of these here and the relevant papers? L16L: "efficient" do you mean effective and if so is the neglect of turbulent momentum loss not part of the reason why these parameters are effective

We mean "effective" and will change it accordingly.

P6836 L1-2: This sentence is difficult to read. P6838 L19-20: The PEST acronym should be defined when it is first introduced in the introduction. P6839 L3-8: I don't understand this sentence. It implies the model is both difficult and easy to calibrate without sufficient detail to explain why this is.

We meant that a full spatial (and temporal) distribution of roughness parameters is easy and straightforward to implement in 2D hydraulic models. Easy is model setup, difficult is the calibration. We will make this more clearly.

P6839 L26: Could you add some clarification about what each of the terms in Eq.4 are. For example, is Q the inverse of the measurement error covariance and does this imply that measurement uncertainty could be considered by the algorithm. I'm not familiar with the approach used here so it would be nice to have some additional clarification of these points.

Cf. reply to Matt Horritt comment. In principle the correlation between observation could be considered, but in the current implementation in PEST is cannot. We will add some more details on the Levenberg-Marqquardt algorithms in the revised manuscript, but since this is a standard optimisation routine explained in a number of textbooks and the manual, we will keep it short.

P6840 L7-8: Given a different objective function can this method be used with time series data (e.g. gauge data), has this been done in other application areas?

Of course, the method can be used with time series data. In fact this is the traditional way of using it. However, the implementation in PEST does not allow usage of a different objective function. If a different function as the squared sum of weighted residuals should be used, the only way is to manipulate the simulation and observation data outside of PEST and have them summed and squared in PEST. A similar approach has to be followed if e.g. inundation extends should also included in the calibration.

P6841 L4: Presumably, PEST runs several simulations with different parameter vectors at the same time as a batch of jobs? If so this should be distinguished from the case where a single simulation runs in parallel (thus quicker) on multiple cores.

Yes, the parallel version of PEST works just like this. We will add the notion that the actual model is not parallelised to avoid confusion.

L13: "if the covariance matrix has been calculated" Is this not always done or is it computationally expensive? Again I'm not familiar with the method so this may be a misunderstanding on my part.

The covariance matrix can not be calculated if e.g. JtQJ of equation 4 can not be inverted (Doherty, 2004), however this was not our case.

P6842 L7-8: Is the gauge upstream or downstream of the site and how far away is it? Presumably the gauge is on the Mulde? How was the flow on the Muhlgraben defined?

The gauge Golzern is about 20 km upstream of Eilenburg on the Mulde. Because also

the channel was modelled two-dimensionally it was not necessary to define the flow for Muhlgraben. The junction of the Mulde and Mühlgraben was modelled explicitly based on bathymetric and topographic data.

L15-18: What data were used to define the DEM?

Please see the answer to M. Horritt and M. Wilson comments.

P6843 L4: "ensemble average roughness" what does this mean?

In each region one roughness coefficient has been defined uniformly distributed. We will rephrase this to avoid misunderstandings.

P6844 L11-14: Could the high roughness be the result of other factors (such as flow errors) or is the model insensitive to channel friction at high flow?

Of course, uncertainty arises from data used and from the same model and we thing the "effective" roughness parameters should somehow account for them. The most likely reason is the definition of the lower boundary, cf. reply to M. Horritt comment.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., 6, 6833, 2009.