**Interactive response to Anonymous Referee #1 (Received and published: 6 May 2015) on "Identification of spatial and temporal contributions of rainfalls to flash floods using neural network modelling: case study on the Lez Basin (Southern France)" by T. Darras et al.**

Authors first want to thank very much the reviewer for his (her) accurate and interesting work. This will surely improve significantly the paper. To facilitate the link with the questions or thought suggested, we propose to respond to, (or comment) each point just after the text of the reviewer. In order to facilitate the reading, responses will be written in blue. Figures were added. For these reasons the interactive response is provided apart, in pdf, as supplement to the author comment.

*1) This paper focuses on the application of the KnoX modelling methodology to extract knowledge about the contribution and timing of different geographical aquifer zones to flash floods in SE France from artificial neural network models. It is a very worthwhile exercise - the karst aquifers of this region are complex and difficult to model physically. Consequently a knowledge extraction approach using data-driven modelling techniques is a sensible and novel solution. It is also excellent to see an example of the use of ANNs for geographical knowledge extraction, rather than the more commonly attempted (and rather uninteresting) lumped catchment 'curve-fitting' tasks. However, I am concerned that the KnoX methodology (which is essentially a method for assessing model input contributions to the output based on the network weights) is not particularly innovative - network weights have been used for more than a decade to understand the importance of inputs (Olden and Jackson, 2002; Kingston et al., 2003; 2005).*

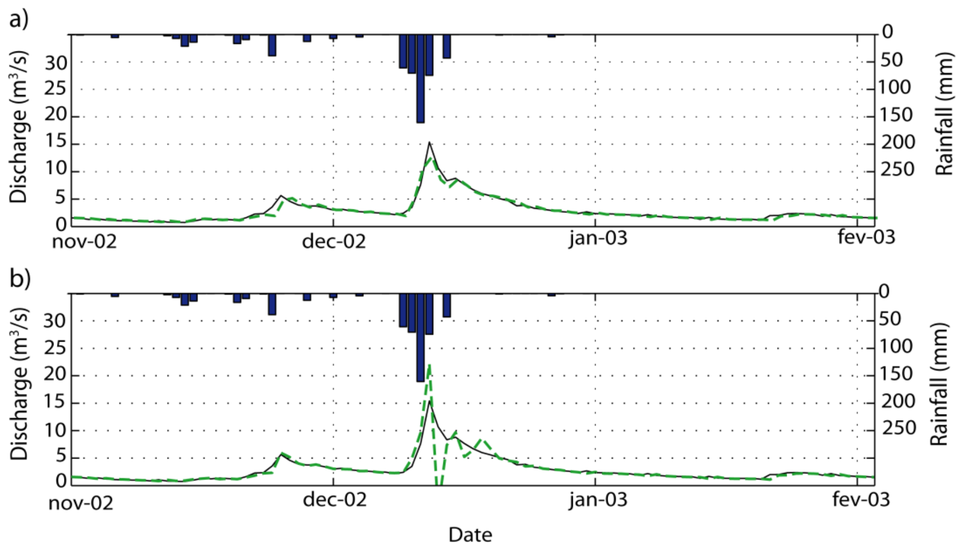
The KNOX method was initially described in the paper (Kong-A-Siou et al, 2013). In this former paper we explained better the origin of the method, without claiming that the idea to consider parameters (or weights) was novel; KNOX was based on the work of Yacoub devoted to variable selection (*Yacoub, B., Bennani, Y., 1997. A heuristic for variable selection in multilayer artificial neural network classifier. Intelligent Engineering Systems Through Artificial Neural Networks 7, 527–532*), which was cited in the previous paper. The main difference with the papers cited previously (Olden and Jackson, 2002; Kingston 2006 (we didn't found Kinston 2003)) is that KNOX is deterministic thanks to the suppression of the random effect due to the initialization of the model. Also the method is different from previous ones: previous ones analyzed variable importance thanks to the strength of the parameters (or weights) aided by sensitivity analysis. In opposition, KNOX it is based on a 2 steps process: first describe the watershed in a block diagram (representing the postulated physical model); Parameters of the model are thus constrained to provide the targeted physical information. Second, each box is implemented using a multilayer perceptron or a unique linear neuron. After that, the contribution of variables can be analyzed and knowledge extracted. In other words, KNOX first constrains the architecture toward the postulated physical model, and second quantifies the physical process. To the best of our knowledge, this two steps procedure is novel. Moreover KNOX was proved (using a virtual artificial aquifer), in the former paper, as sufficiently accurate to extract hydrological knowledge. KNOX method is qualified in L17 as "recent methodology", and the goal of the present paper is "To assess the interest of this methodology" (L20) on another case study. Due to the complexity of both karst aquifer and neural networks, we think that it would be interesting to apply this method to a lot of various configurations in order to assess its efficiency and limitations.

*2) What troubles me more, however, is the fact that the KnoX method has been applied to this aquifer before by this authorship team or members thereof (see several citations of work by Kong-A-Siou et al.). The authors make some reference to this in Section 3.5. and in Section 4 cite that the difference with this paper is the hourly data used to drive the model. This leaves me wondering what the contribution to knowledge is in this paper compared to the several other papers by the same (or similar) authorship team. If this is a repeat of previously published work that is largely the same except for a different temporal resolution of input data, then it feels like only a minor contribution to the literature. Therefore, before it can be accepted for publication, I think that the authors need to be very clear about how this paper develops the other papers by Kong- A-Siou et al., 2013 on the Lez Basin, the new findings / insights that result from this paper, and their relevance and importance for hydrologists.*

We agree with this questioning, which is fully the role of the referee. As it is impossible to rewrite in a new paper all the work made before and published elsewhere, it is usual to cite previous papers.

- In Kong-A-Siou (2011) we applied NN to perform forecast at Lez spring and validate cross validation as a useful method to select the complexity of the NN.
- In Kong-A-Siou (2012) we focused on regularization methods used to select the best complexity (early stopping and weight decay) and we presented one exemplar figure of what is overfitting due to too high

complexity (too many hidden neurons). This figure is the following (top 4 hidden neurons, bottom: 8 hidden neurons). We concluded also that early stopping was efficient.




- In Kong-A-Siou 2014 we compared reservoir model (VENSIM) and recurrent NN to achieve forecasting of discharge and drawdown of water for floods and low water levels. NN was shown efficient for extreme events and VENSIM efficient for intermediate events.

- In Kong-A-Siou 2013 we proposed the KNOX method, explain it, and validate it on a fictitious aquifer. The validation was based on the contributions and time delays of rainfall estimated on 4 zones of the aquifer. The knowledge extracted from the model allows us to have a new vision of the behavior of the aquifer. The study was done on data of the **Lez spring at daily time steps**. Information was useful for water resource (where does the water come from?), and recharge of the aquifer. Only groundwater was considered.

- In the present paper the same basin is considered, with the same delimitation in 4 zones, nevertheless the gauge station is not the same than in 2013: it is the ***Lavalette station*** at the entrance of Montpellier. Also the database is **hourly sampled and includes only flash flood events**. **Addressed processes are thus different**. As the behavior of the basin is very complicated due to surface and underground water floods, and because it is impossible to make measurement during the event (to protect human lives) the application of KNOX method seems interesting to do. After the application of the KNOX method, we found that it was possible to access to a better quantification of processes acting during the flood. We think that this result is really interesting for karst hydrologist as it seems possible to distinguish surface flood and underground flood, thanks to the model as shown in Figure 5. The comparison of contributions and time delays provided in Table 5 allows comparing two different processes acting simultaneously in the aquifer. For us this is interesting, very useful, and we hope that other people redo the same work on other aquifers to explore the potential of this method.

*3) The introduction / literature review is generally well structured and provides a fairly comprehensive and critical overview of the key literature and the arguments from adopting the method used. There are far more examples of the use of the multi-layer perceptron than the two articles cited - a more extensive tabulation would make the review more complete. Similarly, the application of ANN-based models in a spatially discretized structure to deal with heterogenous and complex hydrological behaviour has also been explored before (e.g. with rainfall-runoff models) and it might be worth mentioning these for completeness (e.g. Tsai et al. 2014, Hyd. Proc., 28(3), 1055). The issue of how to select the 'best' or 'correct' input data sets to the ANN is skipped over a little. The authors might like to consider mentioning the sorts of information-based methods that have emerged for selecting model inputs over the last few years (e.g. the Gamma and Evans tests) and explain their choice of inputs a little more thoroughly in light of these ideas.*

We agree that this presentation could be extended. Our opinion is that a scientific paper must be concise (if possible). The paper has a target, and only the necessary literature, useful to understand and enlighten the methodology and results must be cited. This choice implied, in this case, citing 65 papers. We think it is sufficient but this point can be discussed: it is an editorial choice.

*4) I also note that the authors identify ANNs as 'statistical' models. This is, of course, true. However, the term 'data-driven' is perhaps more commonly used to describe ANNs and the authors may wish to alter their terminology.*

Yes it could be possible. Nevertheless we observed that the statistical framework is more and more used in the neural network research field, so in our perception (which can be false) "data-driven" sounds a little bit old.

*5) Section 2 deals with the basic concepts of ANN design and development. The MLP ANN is chosen, but there is not any real justification for this presented. Why not a RBF ANN or some other variant? I think a stronger justification for the MLP would be useful here.*

MLP was chosen (P3686, L3) because of "its properties of universal approximation and parsimony". We thought it was sufficient as well as a lot of authors (also Olden and Jackson, 2002, cited in earlier comments). Moreover, KNOX method was validated in the framework of multilayer perceptron.

*6) The terminology 'stop set' is not standard, but I do think it is clear. 'Overtraining' is more commonly referred to as 'overfitting' and this is a term that the authors might like to adjust. The authors do not explain that the issue of overfitting is exacerbated by data splits that are not fully representative of the signals in the data.*

Yes we agree.

*7) There is a huge literature around methods for achieving representative data splitting to improve the generalisation of ANNs (Holger Maier at the University of Adelaide has published in this area recently) and this literature should at least be cited.*

Yes it is possible; nevertheless the goal of the paper is not a review about data splitting. The goal was not to have the best possible generalization, but a good one taking into account the complexity of the basin and uncertainties on input data (that are considerable in this case). Models T7 and T8 (Table 4) can be viewed as excellent taking into account that they are evaluated on the two highest events of the database. We understand that a meticulous hydrologist want to have the best possible generalization. Nevertheless in the context of Mediterranean events, where peaks of flood are measured with an uncertainty of 20% or 30%, we think that the level of quality of models T7 and T8 is sufficient. For us, the great challenge was to extract knowledge quite similar for all various training set. Regarding this challenge, we thought that results (Table 5 and Figure 5) were surprisingly good.

*8) Section 2.1.2. needs a little work. I really struggled to follow what was going on in the method for identifying the stopping point and had to read the text forensically. I think a flow chart is needed to support the text in 2.1.2 and the authors need to work the text up a little more to improve readability and consistency (e.g. the term 'validation set' creeps in here but this could be confused with the other 'sets' presented earlier). The use of a median value from an ensemble of 50 ANNs to avoid the influence of the random initialisation effect is sensible - but it does risk 'damping' the model outputs. It might be worth being explicit about the impacts that using an ensemble median might have.*

The "validation set" is simply the one of the cross-validation. This information is lacking; it must be added. For readers interested by cross validation, 2 references were provided. Calculation of the median is not used to calculate the prediction; it is only used to extract parameters and knowledge. We agree that the calculation of the median of the output may induce "damping"; in another work we took this effect into account to make the predictions reliable in case of flash floods (Darras et al 2014, Influence of the Initialization of Multilayer Perceptron for Flash Flood Forecasting: Design of a Robust Model).

*9) Section 2.2. deals only briefly with the literature around knowledge extraction. Recent efforts based on partial derivatives have provided useful insights into the physical rationality of ANNs and should probably be mentioned at least (e.g. Mount et al., 2013. HESS, 17, 2827 / Dawson et al., 2014. Jnl Hydroinf. 16(2), 407).*

These two papers are thorough, well written and very honest in describing their aims and limits. They address a very interesting issue: how to validate a model calculated thanks to a calibration process? In our opinion this question must be asked also to practitioners of reservoir models. Indeed it is well known that after calibration, parameters of a reservoir model have generally no realistic values. Despite their high quality, these papers haven't the same goal that KnoX method. Hydrometeorological complexity of the Lez model is incommensurate with the one of the models targeted by these methods. We don't apprehend the significant added value they could provide to the goal of the paper.

*10) The KnoX method description is not particularly easy to follow - the 4 steps in the text should, perhaps, be revised to improve their readability and specificity. For example, step 3 states '...and calculate of the median of the absolute value of each parameter over the ensemble models'. What are these 'parameters'? Are they input values? Are they initialization parameters? Are they values associated with the neurons of the network? This is all very unclear. The KnoX method is central to what follows so the authors really do need to revise this section fully and provide clarity.*

Parameters are the weights of the model. We choose this word in order to be widely understood by the community of neural networks modelers, see for example the book of Dreyfus G. (Neural Networks), and by modelers in hydrology (reservoir models have parameters that are calculated thanks to a calibration phase as well as parameters of the neural model during training). Nevertheless, the reviewer is right, this term is not so clear at this point of the paper. We will include at least the general equation implemented by the multilayer perceptron and the definition of variables and parameters before this description of KnoX method.

*11) Again, a flow chart or schematic diagram could be helpful for all readers here. Section 3.5 - please avoid statements such as 'fed by abundant rainfall'. What is the rainfall - please give measurements. One man's abundance is another's dearth.*

We agree, in Table 1 cumulative rainfall was provided for each event. For example, the maximum is 245 mm in few days.

*12) The legend on Figure 1 is not clear - is the conurbation the hatched area?*

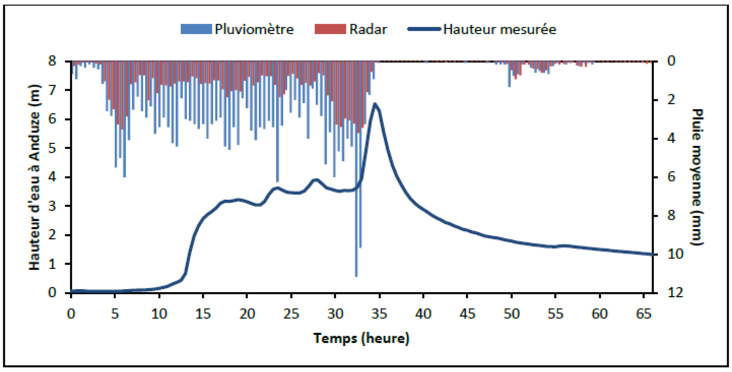
Yes it is, in the original Figure it was better readable.

*13) Section 4.1.1 presents the 'postulated model'. I find it somewhat unsatisfying and poorly argued. It relies on the author's previous papers but little evidence is offered to substantiate the spatial discretisation in the text.*

We were afraid to increase the length of the text, and proposed to the reader to read (Kong-A-Siou 2013) where this issue was widely explained.

*14) The inputs to the ANNs are simply the mean rainfall values in each of the four zones - as determined by Theissen polygons. This is a rather simplistic method for assigning rainfall inputs and does not account for the spatio-temporal heterogeneity of rainfall in the catchment. Is this a potential issue - I would imagine heterogeneity could be high in this catchment? RADAR-based rainfall data might be able to help answer this. What might the impact of such a simple assignment of model inputs be on the final model?*

The availability of rainfall spatially and correctly discretized in time is an actual and important issue. These data are not available and RADAR data suffers in this region from several problems: (1) consistency from one event to another, (2) underestimation of rainfall. They are thus not better than rain gauges even if research in this field is very active to improve RADAR rainfall. Don't forget also that radar data are usually calibrated thanks to rain gauges. Personally we demonstrated empirically in a place near the Lez Basin (Gardon d'Anduze) that RADAR data underestimate rainfall by a mean factor of 25% (with important variability) compared to rain gauge data (by comparing only the rain gauge and the pixel of radar corresponding to the rain gauge); see for example the following figure for a cevenol flash flood at Anduze (sorry, it is in French; "pluviomètre" means rain gauge, and "pluie" means rainfall). Thus using rain gauge data in this basin appeared as more robust.




*15) Simply stating that you 'consider the rainfall information sufficient to carry out this study' (Pg 3695, Line 10) doesn't feel an adequate justification to me.*

As explained before, rainfall data at hourly time step in the south-east zone doesn't exist anymore. Concerned area (in the south) is impervious; therefore there is no doubt about the influence of karst. Karst contains water table under the impervious layer but has no role on flood. Our questioning is linked to the NE and SO zones; indeed the NO zone (full karst area, and surface water flowing to another river (Vidourle)), seems too far from Lavalette station to be able to contribute to underground Flash Flood. The goal of this study is also to better understand the behavior of the basin in order to develop well suited monitoring strategy (p3682 L16-17).

*16) Section 4.2. The authors introduce the term 'window-width' in this section and it appears again in Table 3. I simply don't know what this is - I don't recall having seen it in the text before. Similarly, the authors appear to have experimented with developing models using various numbers of hidden neurons - but I don't recall this important process (the model complexity has a major influence on overfitting propensity) being presented in the text earlier. This leaves me rather confused and of the opinion that the methodological descriptions presented earlier in the paper have not been sufficiently clear or detailed enough.*

Yes, this information is important. Hidden neurons number is determined by cross-validation as indicated p3687 L4-5 and p3695 L16 and Table 3. Window widths refer to the sliding windows of rainfall vectors and of previous discharge vector. These sliding windows are drawn in Fig. 3; their range of investigation and chosen values are provided in Table 3.

*17) I simply can not read Figure 3 and this makes it very difficult to understand the ANN structure that has been used. Similarly, the model outputs in Figure 4 are too small to be useful - I can't see the hydrographs properly. I think that considerably more work is needed here to ensure that the methods and model structure are properly and fully described in the paper and that the model outputs that are being used to validate the model are adequately disclosed.*

We are sorry, original figures were designed with illustrator. They are good in original png format. They are inserted in the following response. The quality is slightly better when clicking on "printer-friendly Version" and zooming (in the internet HESSD discussion).

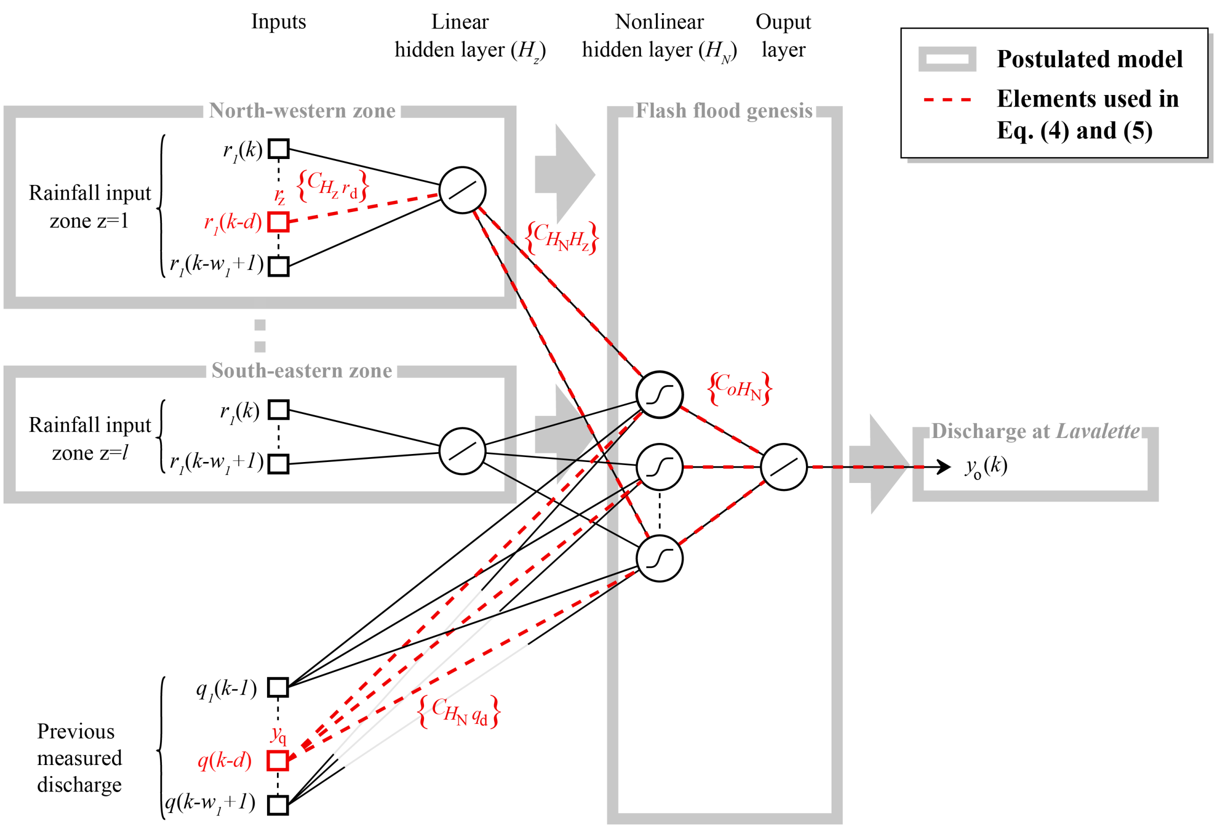



Fig3 Postulated model: grey block-diagram. Three layers multilayer perceptron with linear hidden layer between rainfall inputs and nonlinear layer. Parameters used in Eq. (4) are denoted in red.

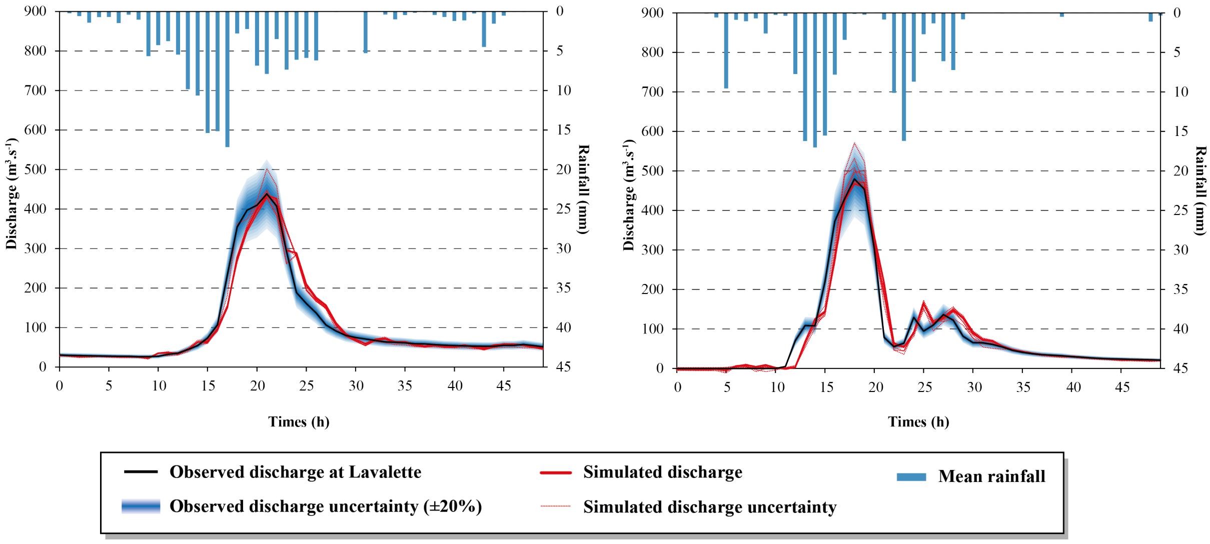



Fig 4. Hydrographs of major events in the database: events 7 and 8. Simulated discharge is the median of outputs coming from the 50 run models (differing by their initialization parameters). Uncertainty on the observed value is the measurement ±20 %. Uncertainty on the simulated value is represented by simulations coming from the 50 run models (differing by their parameters initialization).

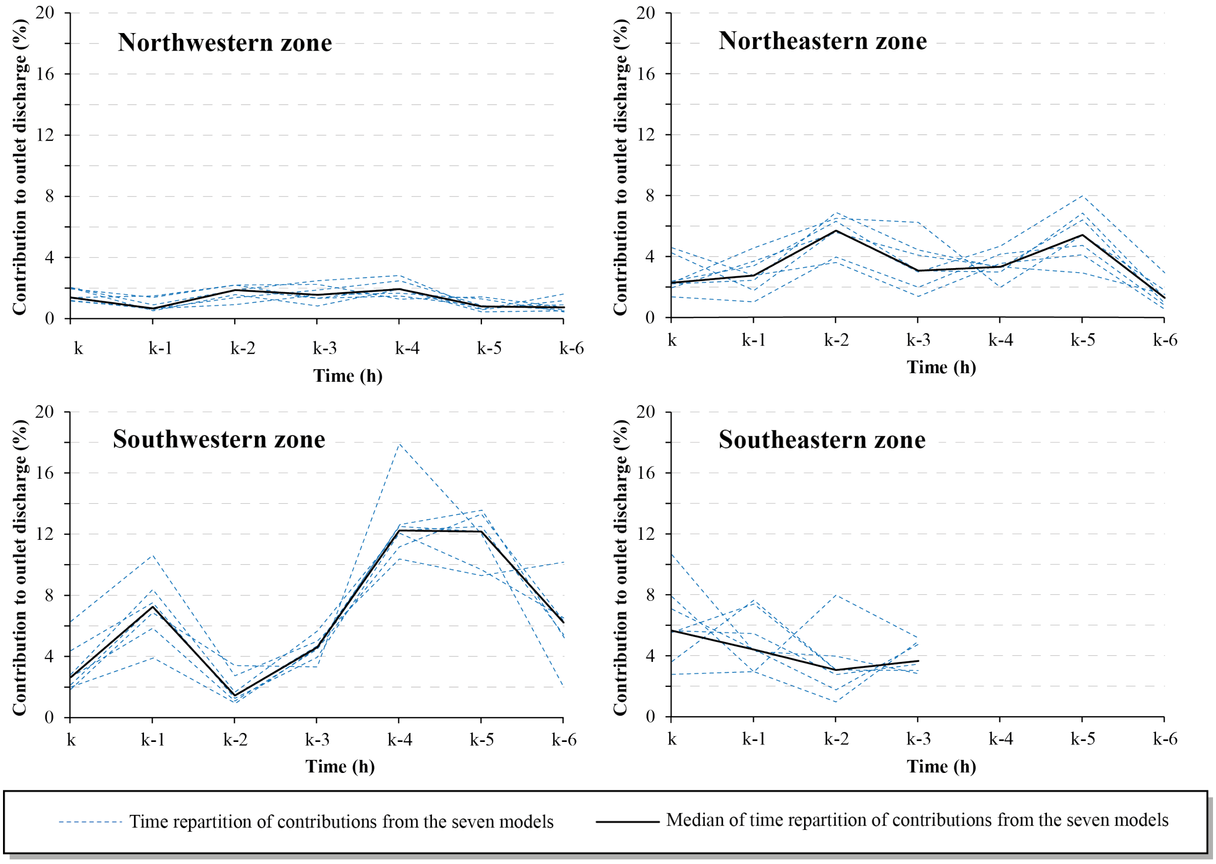



Figure 5. Median and total spread of time distributions of North-western, North-eastern, South-western and South-eastern rainfall inputs contributions calculated from parameters of the 7 designed models

*18) In section 4.3 the nature of the KnoX method becomes clearer, along with what the authors meant by 'parameters' earlier in the paper (they are the network weights). Knox is revealed as a method for determining the influence of each input, on the output, at each time step, based upon chaining of the network weights. The use of network weights to explore and quantify the contributions of different inputs is not particularly new. Work by Olden* *and Jackson (2002) and Kingston et al., (2003, 2006) (which has not been cited) is highly relevant because they did something rather similar. How does KnoX differ from this?*

KnoX differs from these methods in several ways: (1) the first step is to **constrain the model** using the postulated model. This step **diminishes the number** of parameters and excludes the ones physically impossible. The postulated model **guides the simulation model towards the physical solution** and diminishes the equifinality. Thus the box is no longer totally "black". This proposition is different from the proposition investigated in papers proposed by the reviewer (*Olden and Jackson, 2002; Kingston et al.*) because the constraints are not imposed at the level of the parameters (for example make the parameter linking the evapotranspiration to the hidden neurons positive), but at the level of processes in the case of this paper. Using the block diagram of the postulated model, one says that some processes are possible; others are not. (2) The method takes the median of absolute value of each parameter for 50 different initializations in order to be **independent from one specific initialization**. The sign of the parameter is not important as the product of two negative parameters is positive in the chain of parameters product; for this reason and in order to take profit of the "black box" capabilities of ANN, we didn't want to constrain parameters. (3) Thanks to the postulated model it is possible to **extract temporal information** (Fig. 5) and not only variable contributions. This point is fundamental and allows in the present paper to **characterize the hydrological processes**. (4) KnoX method was validated on a fictitious nonlinear aquifer using actual rainfall data.

These elements are described in detail in the previous paper (Kong-A-Siou, 2013). As it contains 32 pages in Journal of Hydrology, it is not possible to re-write it in the present paper. Nevertheless it is necessary to be red in order to fully understand the contribution of the present work.

The papers of Jackson (2002) and Kingston (2006) present a method which computes a chained product of weights, so called "Connection Weight Approach", this approach was not novel in 2002 and this was not a problem, because other interesting works were done thanks to this CWA. It is the case of KnoX method which proposes a way to smartly perform "pruning" and deterministic knowledge extraction. It can be noted that the CWA approach used in KnoX method was inspired by Yacoub and Bennani, 1997, cited in the previous paper (Kong-A-Siou, 2013).

Nevertheless we would be very interested if we can have another reference about a method able to indicate, for example that there is 3 hours delay between underground flood and surface flood, without measurement of each kind of flood. KnoX provided this information in the present paper.

*19) Similarly, the quantification of the partial derivatives of MLPs (Mount et al., 2013 and Dawson et al., 2014 - see earlier citations in this report) are arguably more comprehensive methods for understanding the strength and pattern of influence of model inputs on the output response of an ANN. Again, why is KnoX a preferable method?*

The response was provided above. Even if cited methods seem more accurate than KnoX to analyze the information provided by each parameter (KnoX only calculates the median). It seems to us that they haven't extracted temporal information on a so complex model (26 input variables highly correlated because they are time-shifted rainfall values, and 5 hidden non-linear neurons).

*20) The discussion is simply a summary of the findings of the modelling. This section needs further development to contextualise the KnoX method, its value in hydrological modelling and how it contributes to the range of knowledge extraction methods that have been applied in ANN modelling (see my earlier comments). Moreover, it would be helpful for the readership of HESS to have the contribution of this paper more clearly explained. To help with this, the authors might like to refer to Abrahart et al (2012) (Two decades of anarchy?, Progress in Phys Geog, 36(4), 480) to position their work within the framework set out therein.*

Well, we are happy to note that one of the references cited in our paper is also appreciate by the referee. Clearly the contribution is related to the theme 7: "physical interpretation" as it proposes an integrated and simple method to constrain to network (at the level of the whole architecture), and interpret the parameters to extract physical information.

Thank you one more time for the accurate reading of the paper.