Author's response to reviews

Title: Spatiotemporal clustering, climate periodicity, and social-ecological risk factors for dengue during an outbreak in Machala, Ecuador, in 2010

Authors:

Anna M. Stewart Ibarra (amstew01@gmail.com)
Angel G. Muñoz (agmunoz@iri.columbia.edu)
Sadie J. Ryan (siryan@esf.edu)
Efrain Beltran Ayala (felixbeltran57@hotmail.com)
Mercy J. Borbor (meborbor@espol.edu.ec)
Julia L. Finkelstein (jfinkelstein@cornell.edu)
Raul Mejia (rmejia@inamhi.gob.ec)
Tania Ordoñez (taniaordonez63@hotmail.com)
Cristina Recalde Coronel (gcrecalde@gmail.com)
Keytia Rivero (keytiar@gmail.com)

Version: 6 Date: 21 October 2014

Author's response to reviews: see over
Editorial Office, BMC Infectious Disease

October 21, 2014

Dear Sirs and/or Madams:

Please find enclosed the revised manuscript: “Spatiotemporal clustering climate periodicity and social-ecological risk factors for dengue during an outbreak in Machala Ecuador in 2010,” by Anna M. Stewart Ibarra Angel G. Muñoz Sadie J. Ryan Mercy J. Borbor Efrain Beltran Ayala Julia L. Finkelstein Raul Mejia Tania Ordoñez Cristina Recalde Coronel and Keytia Rivero, for consideration as a Research Article in BMC Infectious Disease. Thank you for handling the paper and thanks to the two reviewers for their kind and helpful comments, which allowed us to improve the clarity of the paper.

I believe that these findings presented in this paper will be of interest to the readership of BMC Infectious Disease, as dengue fever is a growing public health concern and this study presents key findings on community perceptions of dengue be used to develop targeted public health interventions.

The material presented here is original and has not already been published. I confirm that this manuscript has not and will not be submitted for publication elsewhere so long as it is under consideration by the BMC Infectious Disease. I and my co-authors report no conflicts of interest.

I look forward to hearing from you at your earliest convenience.

Yours sincerely,

[Signature]

Dr. Anna M. Stewart Ibarra, PhD, MPA
Research Assistant Professor
Department of Microbiology & Immunology
Center for Global Health and Translational Science
SUNY Upstate Medical University
Response to reviewers

MS: 7511706681324627
Spatiotemporal clustering climate periodicity and social-ecological risk factors for dengue during an outbreak in Machala Ecuador in 2010

We thank the two reviewers for their kind comments and suggestions to improve the manuscript, and to the editor for reviewing these comments and handling the paper. Based on the reviews, we have re-worked our description of the results, particularly in describing the wavelet (climate) analysis. In addition, we have updated the tables to both tidy up the parameter descriptions in the narrative, as reviewer 2 suggested, and to clarify our choices in the analyses of the models (also at the prompting of reviewer 2). We believe this improved clarity and flow. In addition to these points, we addressed reviewer comments individually.

Please the see the responses in red below

Reviewer 1

Abstract> Methods be consistent with the numbers 1,673 vs 1673
We apologize for this error and have corrected the text.

Background
I am a bit confused: are you doing predictions or trying to explain dengue cases (outcome) according to weather, entomological, social ecological variables? please clarify and correct if needed. If here you are explaining; not predicting, then you can't say that "this is the first of such studies conducted in Ecuador" because it is not what you are aiming/doing.
Thank you for this observation. We have removed this phrase, and rewritten the introduction to improve clarity and to highlight how this study builds on our prior work in Ecuador (See 7.15).

Data sources
you reported in the abstract 1,673 dengue cases georeferenced but here in data source you report 1,674. Please clarify.
Thank you for this observation. The correct number is 1674. This has been corrected.

also clarify in the same paragraph if it was monthly, yearly weekly dengue cases?
The data sources in the abstract and the have been rewritten to clarify the data used in the analyses (See 9.8, Data sources).

in section social-ecological risk factors: You refer to Table S1, I am not convinced we need a supplementary info in Spanish. Please remove the table if you agree.
We feel that this table provides important information, because it provides replicable methods (recoding of variables) for others who will work with data from the Ecuadorian census.

in section entomology data I would propose Entomological data instead of entomology
Thank you, we have made this revision.
You found a Pearson Correlation r=0.2 and stated 'strongly associated'; 0.2 is not really strong. Please rephrase.

We found that the Breteau Index (BI) during the first 2 quarters of the year was the vector index most strongly associated with dengue (compared to the other indices). We have clarified this statement to read,

“The average BI during the first two quarters of 2010 (January to June), was the vector index that was most strongly associated with dengue presence (1) or absence (0) (Pearson correlation, r = 0.2, p = 0.0012),” (11-2)

in the same sentence 'accordingly,...multivariate model (BI_12)” this is the name of the variable not of the multivariable model. Clarify or modify the sentence as it is a bit confusing. Ultimately refer to Table 1.

We agree that this was unclear, and have removed (BI_12) from the end of the sentence and referred the reader to Table 1. We have updated our descriptions of model variables in response to this and other reviewer comments, and feel this improves the manuscript considerably.

In climate information it might be worth writing entirely metres above sea level instead of m.a.s.l.

Thank you, we have made this revision.

In statistical analysis, wavelet analysis
Interpolation from monthly dengue cases to weekly dengue cases? that is acceptable but if it is from yearly dengue cases, it is certainly underestimating seasonality. Please explain.

We have improved the paragraph so that is more clearly explained now. No interpolation of dengue cases was done, but of population. This was done in order to compute incidence (14-6)

Same paragraph line 18; “Dengue incidence and rainfall time series’ had non normal probability density functions, thus they were percentile transformed.” do you have any ref to illustrate the method and reasons?

A reference (already cited in the paper) was added at the end of this sentence (14-10).

in Results page 15 line 4 ‘more likely to report ’instead of ’more likely report’

Thank you, we have made this revision.

page 16 line 2 ‘as’ to remove?

Thank you, we have made this revision.

page 16 line 16 : I would add ‘transmission’ to "the climate variables with respect to dengue"

Thank you, we have made this revision.

In discussion
check examples of research articles published in BMC Infectious Diseases, they don't have sections in Discussion

I advise removing the bold titles
Since we are presenting different types of analyses, we feel that it will help the reader if we use similar subheadings in the results and discussion. Here is an example of a recent paper in BMC ID that used subheadings: http://www.biomedcentral.com/1471-2334/14/397

In Discussion->Spatial Dynamics line 12
I understand it is important for you to show you have a paper in review on this topic, but check the BMC ID author guides. "Only articles, datasets, clinical trial registration records and abstracts that have been published or are in press, or are available through public e-print/preprint servers, may be cited" "Any in press articles cited within the references and necessary for the reviewers' assessment of the manuscript should be made available if requested by the editorial office." I would advise to remove the reference or add the reference if it has been accepted.

We have included an alternative reference.

same paragraph line 13, add a point at the end.
page 21, line 11, point to add at the end of the sentence
In References
could you homogeneise your title of journal. Some are with upper case for each word and some title don't
in reference 18, Aedes aegypti to put in italic
All of the previous revisions listed above have been made. Thank you for the observations.

**Reviewer's report:**
The authors present some interesting analyses of spatial and temporal dengue case data in a small city in Ecuador. I like what they've done and recommend it be published, though I have some suggestions for how they could improve the work.

Minor
2.6 Authors listed alphabetically: alas, it seems that something went wrong with the ordering, for MJB comes before EBA.
Thank you, we have made this revision.

6.para1 I think this could do with some references to support some of the claims made.
Thanks for this recommendation. We have added appropriate references throughout this paragraph.

6.16 'An EWS', and henceforth
Thank you, we have made this revision.

6.23 GIS 'based' maps. Generally there are quite a lot of little typos like this in the m/s. A proof read by a native anglophone might pick these up.
The phrase “GIS base maps” is common terminology in GIS, referring to the base map layers that are created and input into spatial models.

7.23 'skill' is almost certainly the wrong word!
Skill is technical word that refers to a statistical property of climate and other forecasts; based on our experience, skill is the appropriate terminology in this context. To clarify, we added the phrase “(predictive ability)” at the end of the sentence. See reference [21] and others therein. In particular that paper shows that the skill level in seasonal forecasts is high enough so early warning systems could provide useful information for decision-making.

9.2 Better to present 95%CIs rather than SEs. Saves the reader a little mental arithmetic.
Thank you, we have made this revision.
10.1-3 Readers won't be able to reproduce this without clarification. We appreciate this comment and have clarified the text and added Equation 1 (10.5-12).

10.13 This is a little unorthodox and might need a stronger justification. We have moved this explanation to page 13 where we discuss model outcome tests, as it is unexpected in the data description, which we appreciate the reviewer calling attention to. We have now included the justification (VIF inflation, model overdispersion), and described the interpretation of that residual in the model 13.3-9).

Note that Table S3 is now Table S2, which identifies the top ranked models. Table S2 from the previous version of the manuscript was dropped, since it showed exploratory analyses of the relationships among the suites of variables, which were not relevant for the final analysis.

11.12 Surely you can only test whether they were random and try to show they are not random? I don't see how you can get a sampling distribution under the hypothesis that they are non-random. We have corrected this statement to indicate that we were testing whether the cases were randomly distributed (11.17).

12.1 'each hypothesis was represented as a suite of variables'---I don't understand this at all, sorry. An hypothesis is usually something like 'this variable is not associated with the outcome', or 'none of this collection of variables is associated with the outcome'. How can a suite of variables by itself be an hypothesis? We apologize for the lack of clarity, and have described this as suites of variables comprising the factors, with ensembles of variables representing those hypotheses. In informatic approaches such as this, model selection occurs on models of suites of variables, which are posed as hypotheses for explaining the data using the least complex model to capture the highest information (12.6-9).

12.7-8 There is no need to put the "(GLM...logit)" part in. You've already said it's a logistic regression. We appreciate the reviewer comment, and have instead taken out the 'logistic regression' part of this, as we wish the methods to be as transparent to the reader as possible (12.13).

12.12 Since you don't present AICs, I think you can drop line 12, and then simplify line 13. I'm pretty bad at algebra but what I got was \( \text{AICc} = \frac{2nk}{n-k-1} - 2\log(L) \). By the way, L should really be the maximum of L over the parameter space. It is the optimum of the log likelihood function, not the entire function, which goes into AIC and friends. While we don't present AICs (we present AICcs, in Table S2 particularly), it is normal practice to present the progression to small sample size in the describing equation, precisely for the algebraic elegance. However, we apologize for overlooking our definition of the parameters, and have now included that (12.19-20).

13.1-2 Please clarify, what exactly is a 'significant temporal scale' or 'significant period'? We have added an explanation in parentheses (13.15).

13.17 What is the [0,1] doing here?
These values indicated the range of the normalized variable, but we just eliminated for the sake of clarity.

13.21-23 I feel you need to clarify what this actually *means*. It may be ok when you address the earlier point on the same issue. We feel that this is clearer now, since we have added the explanation of “significant periods”.

14.4-9 This really should go in the results or figure captions. Methods should just contain something like 'We identified phase angles between pairs of variables, corresponding to GIVE INTERPRETATION, using HOW YOU DID IT, for different time points and periodicities.' We have added a summary of this information to the figure captions (Figs 6-8); however, we respectfully disagree that this should be removed from the methods, as this section explains key aspects of the methodology (especially the use of AR1, MC and COI).

14.14 'greater than 155 cases' This is an odd choice of phrase/number. What is the significance of 155 cases and why is it interesting that 42 neighbourhoods fell above that threshold? Thank you for pointing this out. We have rewritten this section and report the means, range, and proportion of neighborhoods with zero cases reported (15.6-8).

14.18 I don't think I understand this sentence, sorry. The text describing these results has been updated considerably, and in particular, the description of variables that are in phase and the indications given by the arrows is better explained.

“The arrows represent the relative phase, which is indicative of the lags between the two time series, as determined by frequency and time [51]. The direction of the arrows can be used to quantify the phase relationship: arrows pointing horizontally to the right (left) indicate that the two variables are in (anti-) phase. When the signals of two time series are in phase, their maximum amplitudes occur simultaneously.” (17.21)

In addition, the description of the findings is clearer in the results (18.1).

14.21 You claim that the top ranked model is a 'better fit' than the model with everything in it. This cannot be! The model with everything in it must by definition fit as well or better than a model with fewer things in it. Whether the global model is better or more predictive is a different kettle of fish. Using an informatics approach means that a model that is overparameterized – in this case, the global model, containing all variables – does not fit better. The top-ranked model – that which fits better, by having a lower AICc, representing a trade-off between model fit and model complexity – in this case, contains fewer variables than the global model. Note that the ‘global model’ is defined here as the model with all of our hypothesized variables (12.16), not a mathematical algorithm fit to each data point.

15.9-11 Better not to use computerese like BI_12 or popdens. We have updated all data tables and text to remove abbreviated variable names.

15.12-21 I thought this paragraph is not going to help people without a background in signal processing understand what you've done. If precise interpretations could be provided, it would help. So, a 'strong and significant
signal for the 1-2 year periodic band for dengue incidence’ means what? That if dengue incidence is high now, it will be higher than expected for the next 1-2 years? Or 1-2 years from now it will be high again? How do you reconcile the ‘1-2 year’ or ‘0.5-1 year’ periodic bands which you identify in lines 14 and 15, with the ‘1 year’ and ‘2 year’ bands you talk about in lines 18 and 19? If these refer to the same thing, then shouldn't the intervals be retained? And if you have period bands at 0.5-1 year and 1-2 years then is this the same as 0.5-2 years, or is there some additional nuance to interpretation that I'm missing?

We've added a note explaining this. We have changed to 1 yr and 2 yr bands. The original idea was to directly indicate that there's a range of periods associated with each band (as the figures show clearly) and not a specific monochromatic period. With respect to the last questions: there're definitely two different bands in the spectrum, arguably related to different physical mechanisms (that'll be studied somewhere else). (16.11-19)

16.2-4 Missing main clause here
This has been corrected (17.2-4).

16.4-16 Again I think most readers will need more help with this. Your work is interesting but if the reader can't understand, it's wasted.
Agreed. We have added additional interpretation of the results in those lines (17.7-19)

17.17-9 Question: to what extent are cases in Machala infected at home vs elsewhere in the community, eg at work or school? Is this known (or estimable)?
We are currently conducting field studies to improve our understanding of transmission dynamics, but unfortunately this information is not yet known in this region.

19.5 'most important' in what sense?
This sentence has been clarified as follows: “We found that the combined housing condition variable (HCI and piped water) was the most important risk factor for dengue transmission, as indicated by the magnitude of the best-fit model parameter estimate (Table 2A); this parameter was also a significant variable in all other top models (Table S2).”

19.10 omit first comma
This has been corrected.

20.23-21.6 This part seems out of place. First, I don't think they do actually test hypotheses using AIC or variants. Second, the AIC itself is also arbitrary so it's a bit odd to state that p-values are arbitrary and imply that AIC isn't. In any case, the thresholds for AIC were determined from 'standard' alpha levels from hypothesis testing (a challenge left to the reader is to consider nested hypotheses, for which the LRT and AIC can both be used, and work out where the delta_AIC <2 threshold comes from). I think you could drop this part without any real loss to your paper.
We have taken the suggestion above and simply removed this part of the paper. We agree that the reader can derive the threshold for AIC comparison (2) that has become standard, and can explore the literature to understand where the likelihood camp stands in the frequentist-Bayesian debate. This is really not pertinent to this study, and we do not wish to distract the reader.

21.8 I'm pretty sure the AIC/glmmulti combination also leads to estimates that are biased away from 0. You'd need to introduce some kind of penalty to the
objective function when estimating parameters to get around that---eg with lasso or a prior.
I don't entirely understand what the reviewer is asking with this question; part of the purpose of the glmutli framework is that it avoids the biases in parameter estimations that arise with stepwise selection processes. It simply uses an information criterion (IC, in this case, AICc) to compare all models equally; this prevents dependence on starting state, and there is no need to evoke a shrinkage technique (like lasso).

References: ref 38 is missing a journal. I suggest the authors go through each reference carefully to check for other errors.
Thank you for noticing this. We have reviewed and updated all references.

Table 1: I'd omit the computerese '4pplbedrm' etc---it's no longer the 1980s!
The variable names have been modified, and we have reformatted the variable formats to improve clarity.

Population density is missing a number
This has been corrected.

Table 2: Again, omit computerese. What is the outcome variable here? How to interpret the estimate? Suggest omit intercept, make the column be OR or similar effect size with 95% CI in brackets, omit the Std [SE?], possibly omit VIF unless you think it’s important. Caption needs more detail.
The variable names have been modified to improve clarity, as indicated above. The outcome variable is the presence/absence of dengue in a neighborhood. This has been clarified in the table caption. We have found that for this kind of analysis, the relative effect size can be determined from the parameter estimates and OR values are confusing for the reader. For example, the OR for the variable “Head of household is a woman” = 2373. VIF values are important in order to indicate that there is no important remaining effect of multicollinearity in the model.

Fig 1: change avg to mean to avoid abbreviation? Can some of the figures be merged into more multipanel figures?
Fig 3: what are the dark dashed lines?
We have combined figures 1 and 3, updated the data to reflect the full time series of available data, and have confirmed that all the variables are defined in the legends.

Fig 7-9: I don't understand the years "at all". Years are intervals in time, but they've been assigned to specific points on the x axes. So what are the points in between? Say for 2003 and 2004, half way between these is what?
The “points” in between are weeks, and the written year indicates the start of the corresponding year. This is a fairly common practice when making wavelet plots. To clarify, we have added an extra sentence at the end of Figs 8 and 9 to explain the black boxes (something that we didn’t do before and it’s important to understand the figure), which we think will make clear that there are weeks on each year. In addition, we have added information about the COI and arrows to clarify the figures for the reader.