We thank the Reviewer 1 for his/her insightful comments and suggestions. We have implemented all applicable recommendations that improve the quality of the presented work. Point-by-point responses are given below. In the following, the comments raised by Reviewer 1 are split into parts and copied in bold fonts to facilitate understanding of our answers.

.

REVIEWER #1

The material provided in this paper is of interest for climate change impact studies in catchment scale hydrology. Authors makes use of performance metrics to assess several Ensembles RCMs on their ability to reproduce the precipitation and temperature regime over selected areas. A state of the art reference dataset (EOBS) is used to evaluate the RCMs output. This paper addresses scientific questions relevant to the scope of HESS, and presents novel concepts to a degree. However some of scientific methods that are used in this paper lack of solid scientific basis and there are several points of criticism that make the manuscript inadequate for publication in HESS in the present form. Each point is discussed in detail below:

a)The title of the paper does not reflect the content of the manuscript. The fact that precipitation and temperature are considered for evaluation, does not justify the "hydrological applications" and "catchments" parts of the title. An alternative title should exclude those parts of the title, e.g. "Regional climate models performance in the representation of precipitation and temperature over selected areas". Following the Reviewer's suggestion, we changed the title of the manuscript to read as:

"Regional climate models' performance in representing precipitation and temperature over selected Mediterranean basin areas".

b)Authors compare the RCMs data to the E-OBS dataset on the period 1951-2010. An important point, that is not given the proper significance, is that the ENSEMBLES RCMs are run from 1951(or 1961) to 2000 under the control emission scenario, while the simulations from 2000 on, are under the A1B emission scenario. The comparison of E-OBS to the RCM data between 2000 and 2010 is valid only under the assumption that this decade's emissions followed the A1B scenario.

b) In principle, we agree with Reviewer's #1comment. Two are the reasons why we consider the 60-year period 1951-2010 instead of the 50-year period 1951-2000: i) to account for the largest common period between ENSEMBLES runs and E-OBS data, in order to more robustly estimate the monthly and yearly climatologies, and ii) to produce and intercompare results for different climatological periods of constant length (see Figures 9 and 10). Since the 30-year time frame has been widely used in many of the recent works, it was adopted to facilitate intercomparisons with other studies. Note, moreover, that given the high spread of the results from different models, the ensemble averages and error metrics for the 60-year period 1951-2010 are almost identical to those computed for the 50-year period 1951-2000.

c) Authors introduce performance metrics to rate the overall RCM ability to be used for hydrological impact studies. Firstly, in equations (6) and (7), authors use weighting factors of 50% to account for both P and T. However they do not elaborate with the selection of the specific weight. The deviation of P and T from the observations affect in different degrees the efficiency of a hydrological model, thus the weights are arbitrarily defined. A weighting factor in this case should be subject to the hydrological model used, the climatology of the basin etc. Furthermore, a metric that assesses the overall RCM performance for hydrological applications should consider the ability of the model to reproduce the ET component of the hydrological cycle.

Reviewer #1 is right. Selection of appropriate weighting factors should be conducted on the basis

of the hydrological model used, its representation of the hydrological processes, the local climatology, and the catchment characteristics. However, the results presented in this work have been obtained within the activities of a project (i.e. CLIMB) where a variety of hydrological models has been used. More precisely, in some of the basins, CLIMB partners used both fully- (e.g. tRIBS) and semi-distributed (e.g. SWAT) hydrological models with fundamentally different parameterisations. In this setting, one has two options: i) change the weighting factors used for precipitation and temperature depending on the model and catchment, or ii) maintain a constant weight (e.g. 50%) for both precipitation and temperature independently from the model or catchment. Unless the weighting factors can be accurately determined (i.e. based on some "rules"), the first option would maximize the uncertainty of model evaluations. Since uncertainty plays the most influential role in assessing the relative performance of climate model simulations (see e.g. Figures 9 and 10), we chose the second option, which may introduce some slight bias, but does not significantly affect results. Concerning evapotranspiration (ET), the main issue is the lack of available data. Since E-OBS dataset doesn't include ET observations, obtaining ET estimates based on other variables would introduce additional uncertainties in model evaluations.

To clarify the above issues, we have added the following text in the revised version of the manuscript (page 11, lines 325-331):

"It is worth mentioning that proper weighting factors in Eqs. (6) and (7) should be in principle determined by taking into account the structure and parameterization of the hydrological model used, its sensitivity to different forcing variables, as well as the climatology of the basin. However, within the CLIMB project a variety of hydrological models have been applied, including fully distributed hydrological models as well as semi-distributed models, in a number of catchments with different climatologies. Under this setting, the most neutral option (i.e. a 50% equal weight for both precipitation and temperature) has been chosen."

D)This study makes use of small areas to compare the performance of several RCMs for their ability to reproduce P, T fields. However the performance of the RCMs over such limited in number and extend areas cannot consist a reference for hydrological applications in general. There are papers in the literature that address the questions that this paper tries to address, in a more holistic way over larger domains (see Kjellström,2010).

The present work does not intend to provide a "general reference" for hydrological applications. In our opinion, and as discussed in several parts of the manuscript, the uncertainty associated with the outputs from different climate models does not allow general assessments that are scale independent; see e.g. page 9125, line 25 in the discussion paper. On the contrary, the manuscript describes a general method to be used when facing the practical problem of selecting a proper set of GCM/ RCM combinations to run hydrological models in small catchment areas. This important issue was addressed during the CLIMB project, and it is of interest to a wide audience of hydrologists and practitioners. Moreover, the manuscript neither intends to provide future climate assessments, nor to evaluate the general performance of ENSEMBLES over large areas, as has been effectively done in the cited papers (see e.g. below). This is explicitly stated in page 9125, line 25 of the discussion paper: *" Our study suggests that, when interest is at relatively small spatial scales associated with hydrological catchments, as it is the case of CLIMB project, validation of CM results should be conducted at a single-basin level, rather than at macro-regional scales. In this case, it is necessary to check models' skills in reproducing prescribed observations at specific river basins, since averaging over quite large areas might bias the assessment. For example, for Riu Mannu, Thau and Chiba catchments, model performances can vary significantly (see Sect. 5), even though these catchments are included in the same large-scale area in Christensen and Christensen (2007) study."*

However, as suggested by Reviewer#1, in the revised version of the manuscript, we have added proper reference to the work of Kjellström (2010) (page 14, line 463).

Other comments:

I would like to bring to Authors' attention two publications that elaborate with model comparison techniques. The one is Taylor (2012) who introduces a method to summarize the degree of correspondence between various simulated and observed fields using a single diagram. The second is a performance metric introduced by Perkins (2007) that may be relevant to the "errors at 100 uniformly spaced probability levels" (Page 9124 – line 1 and Figure 8) (This citation is already used in the literature review of this manuscript).

The paper was revised to include proper references to the works of Taylor (2001) and Perkins (2007) (see page 3, lines 54-76).

Figure 1 is vague and does not provide the information described in the figure caption i.e. the location of each considered area for the comparison.

Figure 1 was regenerated to better illustrate the areas used for verification purposes (see Figure 1 in the revised paper). Also, a clarification has been added to the Figure caption.