Response to reviewers of "The GRENE-TEA Model Intercomparison Project (GTMIP): Overview and experiment protocol for Stage 1," by Miyazaki et al.

We greatly appreciate Referee #1 for his/her positive evaluation on the framework, targets and analysis plan of the intercomparison project, and for helpful and constructive comments and suggestions on the discussion paper. Overall, we have substantially reorganized and rephrased the manuscript; based on the valuable comments from the two referees to clearly present that this is such a paper as for "model intercomparison descriptions, including experimental details and project protocols" (GMD's Aims and scope). Below, referee comments are shown in red and italic, and our specific replies follow.

Referee #1

The paper describe the protocol and very preliminary results for the stage 1 of the GRENE-TEA model inter-comparison project. The proposed inter-comparison project is very interesting as it aims to compare very different kinds of models in their ability to simulate both biophysical and biogeochemicals processes of the pan-artic region. These region is obviously of first importance since it should experience the highest temperature change in the future. Because of the permafrost, there is a risk of large feedback with soil and soil carbon and then it is a very vulnerable region. Moreover the models have not been extensively evaluated for these regions. So this is and interesting subject and I am sure that project could lead to very interesting results.

However the project is still at a very early stage. Then my main concern about the paper in it present form is that there is almost no results presented. Then I really not understand why the authors want to publish a paper at this early stage of the project and not waiting the end of the stage1 to present a more complete analysis of the results ? I would eventually understand if the experiment protocol leads to development of specific new *tools. But this is not the case here where the protocol is relatively standard for such kind of experiment. The paper gives a promising analysis plan, looking to the cause of differences between models, studies at different temporal time scale and conducting several sensitivity tests. Then part dedicated to presentation of results is very frustrating as it is less than one page and stay very descriptive without any real discussion about results. For all these reasons I think that the paper cannot be published in its present form and should be resubmitted with a complete analysis of the stage 1 results when, I am sure it will be a very interesting and useful contribution for the modelling community.*

We thank Referee #1 for his/her very positive and encouraging evaluation on this model intercomparison project. We think, however, the publication of the protocol as a separate paper is grounded and reasonable for the following reasons:

1. As stated more clearly in the revised new section "3. Analysis plan and exemplary results" (an amalgamation and expansion of the previous Sections 2.4 and 3), the analyses and the consequent resulting papers will vary in topics, each of which will be authored by a different group of researchers, thereby this protocol paper acts as a guiding paper.
2. This paper also provides an overview of the entire GTMIP activity (including both Stages, 1 and 2), which include the frameworks of the model intercomparison *per se*, as well as the collaborations among modelling and empirical communities in Japan that assisted the activity (cf. new Fig. 1). Creation of driving data (for spin-up and the experiment), and the boundary conditions on ecosystem activities is already a product of such collaborations, and described in the companion paper (Sueyoshi et al., accepted for publication as a discussion paper in Earth System Science Data Discussion (ESSDD)). (new Section 1.2 GRENE-Arctic project and GTMIP)
3. This model intercomparison is a unique project among previous MIPs in terms of its scope (ranging from physical to ecosystem processes) and geographical domain (the Arctic region) of the target (revised Section 1.

Introduction).

4. Publication of the protocol is a part of interactions in the project, and a critical mean for recruiting new participants. Participation in this GTMIP stage 1 is voluntary and open to any interested parties (modellers, groups, and/or institutions). Actually, a new participant joined after the publication of the discussion paper.

Considering those points and issues, we have reorganized and rephrased the manuscript. Major revision has been done in Abstract, and Introduction, which are given below:

- Abstract

"As part of the terrestrial branch of the Japan-funded Arctic Climate Change Research Project (GRENE-TEA), which aims to clarify the role and function of the terrestrial Arctic in the climate system and assess the influence of its changes on a global scale, this model intercomparison project (GTMIP) is deliberatively designed to 1) enhance communication and understanding between the "minds and hands" (i.e., between the modelling and field scientists) and 2) assess the uncertainty and variations stemming from variability in model implementation/design and in model outputs using climatic and historical conditions in the Arctic terrestrial regions. This paper provides an overview of all GTMIP activity, and the experiment protocol of Stage 1, which is site simulations driven by statistically fitted data created using the GRENE-TEA site observations for the last three decades. The target metrics for the model evaluation cover key processes in both physics and biogeochemistry, including energy budgets, snow, permafrost, phenology, and carbon budgets. Exemplary results for distributions of four metrics (annual mean latent heat flux, annual maximum snow depth, gross primary production, and net ecosystem production), and for seasonal transitions are provided to give an outlook of the planned analysis that will delineate the inter-dependence among the key processes, and provide clues for improving model performance."

- 1. Introduction

“The pan-Arctic ecosystem is characterized by low mean temperatures, snow cover, and seasonal frozen ground or permafrost with a large carbon reservoir, covered by various biomes (plant types) ranging from deciduous and evergreen forests to tundra. The Arctic climate and ecosystem differ from the tropical and temperate counterparts primarily because it is a frozen world. Moreover, the terrestrial Arctic varies from area to area according to the location, glacial history, and climatic conditions. However, sites, networks, and opportunities for direct observations are still sparse relative to the warmer regions owing to physical and logistical limitations. To investigate the impact of climate change in this region, a number of studies using both analysis of observed data and numerical modelling have been carried out (e.g., Zhang et al., 2005; Brown and Robinson, 2011; Brutel-Vuilmet et al., 2013; Koven et al., 2011, 2013; Slater and Lawrence, 2013). Various numerical modelling schemes have been developed to treat physical and biogeochemical processes on and below the land surface. Some of these processes are site-specific or process-oriented, while others are implemented as components of atmosphere–ocean coupled global climate models (AOGCMs), or Earth system models (ESMs) to interact with the overlying atmosphere. Among these processes, snowpack, ground freezing/thawing, and carbon exchange are the most relevant and important processes in terrestrial process models (TPM) for investigating the climate and ecosystem of the pan-Arctic region.

1.1 GRENE-Arctic project and GTMIP

The GRENE-TEA model intercomparison project (GTMIP) was originally planned as part of the terrestrial research project of the GRENE Arctic Climate Change Research Project (GRENE-TEA) to achieve the following targets: a) to pass possible improvements regarding physical and biogeochemical processes for Arctic terrestrial modelling (excluding glaciers and ice sheets) in the existing AOGCM terrestrial schemes for the AOGCM research community, and b) to lay the foundations for the development of future-generation Arctic terrestrial models. The project, however, involves groups of researchers from different backgrounds/disciplines (e.g., physics/geophysics, glaciology, biogeochemistry, ecosystem, forestry) with a wide range of research methods (e.g., field observations, remote-sensing, numerical modelling), target domains (e.g., Northern Europe, Siberia, Alaska, Northern Canada) and scales (from site-level to Pan-Arctic). As is often the case, multi-disciplinary opportunities were limited, initially creating a considerable challenge for the project (Fig. 1a). Communications between groups (e.g., modelling and field studies, physical and ecosystem disciplines, process-oriented and large-scale modelling), if any, were inconclusive and sporadic. Observational practices and procedures (e.g., variables to measure, equipment to use, standard zero depth for ground measurements) were different among groups and disciplines, and lacked standardization. Although each individual group had the needs and intention to interact with other groups, the requisite collaboration could not be achieved. Opinions obtained in the early stages revealed hidden quests for possible collaborations for "observational data for driving and/or validating data", "use of numerical models to test empirical hypothesis gained at the field", "interpretation of observed phenomena", and "optimization of observation network strategies." As a result of this situation, the model intercomparison project was deliberately blueprinted to promote communication and understanding between modelling and empirical scientists, and among modellers: the GTMIP protocols and datasets are set to function as a hub for the groups involved in the project (Fig. 1b). It also aimed to enhance the standardization of observation practices among the GRENE-TEA observation sites, and to form a tight collaboration between the field and modelling communities, laying a cornerstone for creating the driving dataset (details of the Stage 1 driving data and their creation as a product of collaboration between modellers and field scientists are documented by Sueyoshi et al. [2015]).

1.2 Model intercomparison for the terrestrial Arctic

Since the 1990s, a number of model intercomparison projects (MIPs) have been carried out, focusing on the performance of TPMs, AOGCMs, and ESMs; examples include PILPS (Project for Intercomparison of Land-Surface

Parameterization Schemes; Henderson-Sellers, 1993), SnowMIP (Snow Models Intercomparison Project; Etchevers et al. 2004; Essery et al. 2009), Potsdam NPP MIP (Potsdam Net Primary Production Model Intercomparison Project; Cramer et al., 1999), C4MIP (Coupled Climate–Carbon Cycle Model Intercomparison Project; Friedlingstein et al. 2006), CMIP5 (Coupled Model Intercomparison Project; Taylor et al. 2012), and MsTMIP (Multi-scale synthesis and Terrestrial Model Intercomparison Project; Huntzinger et al., 2013), to name a few.

For snow dynamics, SnowMIP2 showed a broad variety in the maximum snow accumulation values, particularly at warmer sites and in warmer winters, although the duration of snow cover was relatively well simulated (Essery et al., 2009). The same study also noted that the SnowMIP2 models tend to predict winter soil temperatures that are too low in cold sites and for sites with shallow snow, a discrepancy arguably caused by the remaining uncertainties in ecological and physical processes and the scarcity of winter process measurements for model development and testing in the boreal zone. The CMIP5 models simulated the snow cover extent for most of the Arctic region well, except for the southern realm of the seasonal snow cover area (Brutel-Vulmet et al., 2013). The poor performance of some of the TPMs in this region is due to an incorrect timing of the snow onset, and possibly by an incorrect representation of the annual maximum snow cover fraction (Brutel-Vulmet et al., 2013). For ground freezing/thawing processes, Koven et al. (2013) showed the current status of the performance of AOGCMs for permafrost processes based on CMIP5 experiments. There was large disagreement among modelled soil temperatures, which may have been due to the representation of the thermal connection between the air and the land surface and, in particular, its mediation by snow in winter. Vertical profiles of the mean and amplitude of modelled soil temperatures showed large variations, some of which can be attributed to differences in the physical properties of the modelled soils and coupling between energy and water transfer. This appears to be particularly relevant for the representation of organic layers.

For the biogeochemical cycles, a number of studies based on MIPs have been carried out. The broad global distribution of net primary productivity (NPP) and the relationship of annual NPP to the major climatic variables coincide in most areas with differences among the 17 global terrestrial biogeochemical models that cannot be attributed to the fundamental modelling strategies (Cramer et al., 1999). The ESMs in CMIP5 use the climate and carbon cycle performance metrics, and they showed that the models correctly reproduced the main climatic variables controlling the spatial and temporal characteristics of the carbon cycle (Anav et al., 2013). However, several weaknesses were found in the modeling of the land carbon cycle: for example, the leaf area index is generally overestimated by models compared with remote sensing data (Anav et al., 2013); NPP and terrestrial carbon storage responses to CO2 increases greatly differs among models (Hajima et al., 2014); current ESMs displays large variations for the estimated soil carbon amounts, in particular for northern high latitudinal regions, and lack the capability to represent the potential degradation of frozen carbon in permafrost regions (Todd-Brown et al., 2014). The future projection by ESMs suggests that the carbon sink characteristic will increase in northern high latitudes, although there are some uncertainties, such as nutrient limitations in CO2 fertilization, the effect of soil moisture on decomposition rates, and mechanistic representations of permafrost (Qian et al., 2010; Ahlstrom et al., 2012; Arora et al., 2013). It should be noted that the reference observation data used for these evaluations are prone to uncertainties due to random and bias errors in the measurements themselves, sampling errors, and analysis error, especially for biogeochemical variables such as land gross primary productivity (GPP) (e.g., Anav et al., 2013; Piao et al., 2013). Based on the outcomes of these MIPs, TPMs have improved their performances.

At scales from a continental level (including those mentioned above) to a site level (model-observation comparisons; e.g., Zaehle et al., 2014), different MIPs have also been conducted, and generally study physical or ecosystem processes separately. PILPS (Henderson-Sellers et al., 1993) and a series of snow MIPs (Etchevers et al., 2004; Essery et al., 2009) are well-known MIPs for physical processes, targeting hydrology and snow dynamics. Recently, an

MIP for tundra sites has been conducted, but its focus is limited to soil thermal dynamics (Ekici et al., 2014). In turn, ecosystem MIPs on continental scales have two predecessors: i.e., the North American Carbon Program Site Synthesis (Schwalm et al., 2010) and CarboEastAsia-MIP (Ichii et al., 2013). Although both MIPs employ multiple terrestrial biosphere models to different eddy-covariance measurement sites (Schwalm et al. (2010) with 22 models for 44 sites in North America; Ichii et al. (2013) with 8 models for 26 sites in East Asia), boreal and Arctic sites were not the major targets. In other studies targeting specific eco-climatic regions, the Arctic was again not the main domain: Jung et al. (2007) assessed GPPs for Europe, and Ichii et al. (2010) for Japan. Rawlins et al. (2015) assessed carbon budget differences among several GCM-compatible models in northern Eurasia, with little examination of the physical processes. In other regions than the Arctic, there have been cross-sectional evaluations of physical and ecosystem processes, such as Morales et al. (2005), evaluating carbon and water fluxes in Europe, and de Gonçalves et al. (2013), the LBA-Data Model Intercomparison Project (LBA-DMIP), analysing water and carbon fluxes in the Amazon.

The GTMIP consists of two stages (Fig. 2): one dimensional, historical GRENE-TEA site evaluations for examining the model's behaviour and its uncertainty (Stage 1), and circumpolar evaluations using projected climate change data from GCM outputs (Stage 2). Hereafter, we describe the Stage 1 protocol. This stage aims to evaluate the physical and biogeochemical TPMs through three-decade site simulations driven and validated by the GRENE-TEA site-derived data. It calls for broader participation in the activity from a wider community to assure robust assessments for model-derived uncertainty, and to efficiently investigate the terrestrial system response to climate variability considering the diversity of the pan-Arctic sites. Thus, the scope and geographical domain of GTMIP Stage 1 is unique in its target of the Arctic region, including both taiga and tundra, and in its evaluations of the behaviour of the energy-snow-soil-vegetation subsystem, employing a wide range of models from physical land surface schemes to terrestrial ecosystems. ”

- New Figures 1 and 2:

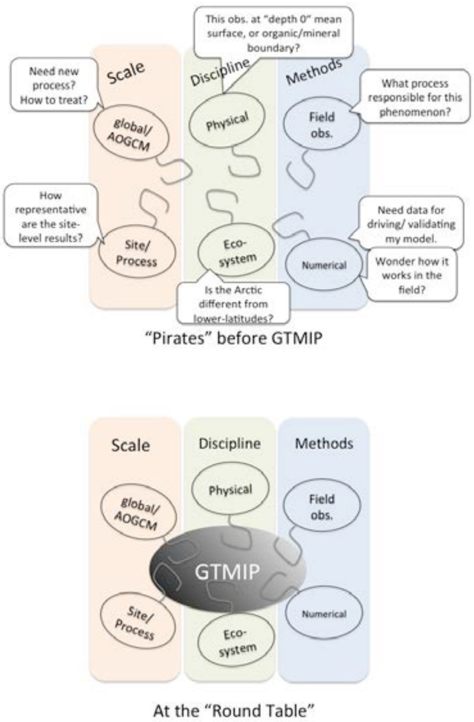

Figure 1. “Pirates of the Arctic” sit together at thea Round Table

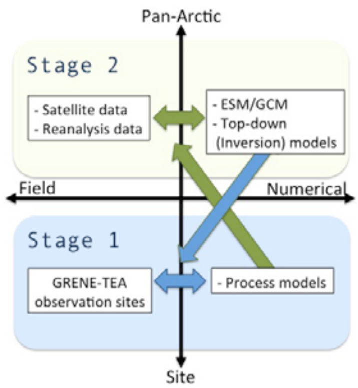

Figure 2. Schematic diagrams for stage 1 and 2 of GTMIP

Since this paper is intended to be a model experiment description paper, the “Preliminary results” section in the original manuscript was meant to provide “sample model output” as “descriptions/figures of model results to give an overview of the project.” In order to addressed the concern of the referee that the “*part dedicated to presentation of results is very frustrating as it is less than one page and stay very descriptive without any real discussion about results*” we have merged the “2.4 Analysis plan” section to the new Section “3. Analysis plan and exemplary results”, and showed more clearly the descriptions of the model results with sample figures for topical analyses (Figures 5-8), and cross-sectional analyses (Figure 9; seasonal transitions).

- 3. Analysis plan and exemplary results

"This section presents the analysis plan for GTMIP Stage 1 and sample outputs based on already submitted materials. To answer the key questions for the target processes proposed in Sect. 2.1, we plan to analyze the model output by describing the model–model and model–observation differences, discerning the cause of these differences, and investigating parameter sensitivity. The outputs of multiple models will be compared in terms of the metrics shown in Table 3. These metrics are divided into five categories (i.e., energy and water budget, snowpack, phenology, subsurface hydrological and thermal states, and carbon budget). For terrestrial climate simulations on the decadal scale, the most important outputs are the latent heat flux (energy and water budget) and the net ecosystem exchange (carbon budget). The latent heat flux (evapotranspiration) is the essential driver of precipitation inland at high latitudes owing to high rates of recycling (e.g., Dirmeyer et al., 2009; Saito et al. 2006). Net ecosystem exchange (NEE) plays a fundamental role in determining global CO_2 concentrations by determining whether a site forms a carbon source or sink (e.g. Abramowitz et al., 2008; Mcguire et al., 2012). NEE represents the net land–atmosphere CO_2 flux, and a positive NEE represents net loss of CO_2 from the land to the atmosphere (i.e., carbon source; Mcguire et al., 2012). Although NEE is commonly used for tower flux observations and some TPMs, the net ecosystem production (NEP) is used in GTMIP for both the observed and simulated values because it is more widely used in non-biogeochemical communities. A positive (negative) value of NEP represents a carbon sink (source).

Analyses will be organized and conducted in the following manner. Topical analyses, constituting major subsets of the project outcomes, will evaluate characteristics of model performances and their inter-site variations within each of the above five categories, while cross-sectional analyses between categories will explore the functionality and strength of interactions between processes. These analyses will be utilized for mining crucial processes to improve the site-level TPMs as well as large-scale GCM/ESM components.

First, the focus will be on model output variability for both the inter-annual and the inter-decadal time scales, based on the output time series over more than 30 years. Inter-site differences will also be evaluated for the four GRENE-TEA sites in the Arctic region, each of which has distinct characteristics. The vegetation type for three of the four sites is forest (two evergreen conifer: FB and KV; one deciduous conifer: YK) and the remaining site is tundra (TK). Three sites (FB, TK, and YK) are in the permafrost region, while KV is underlain by seasonally frozen ground. Figures 5–8 show statistical summary comparisons of the model outputs by site (the land cover and soil type parameters used for the simulations are shown in Table 2), expressing inter-model variations for physical and biogeochemical models using box plots for four variables of the metrics mentioned above: the annual mean latent heat flux (Qle_total_an), the annual maximum snow depth (SnowDepth_max), the annual gross primary production (GPP_an), and the annual net ecosystem production (NEP_an), respectively. When observed values were available (i.e., latent heat flux for FB for 2011–2013 and YK for 1998, 2001, 2003, 2004, 2007, and 2008), they are shown by black dots.

Second, the cause or attributes of the differences among models, or between models and observations, will be explored by employing statistical evaluations such as multivariate analyses and time series analyses on the metrics and individual eco-climate variables. This will improve understanding of the interrelation between the incorporated processes in each model. Figure 9 shows an exemplary comparison of a seasonal transition in the snow-permafrost-vegetation sub-system, expressed similarly by box plots. The figure summarizes the average dates for (from bottom to top) the completion of snow melt, the thawing of the top soil layer, the start and end of greening, the freezing of the top soil layer, and the start of seasonal snow accumulation. A comparison of the timings of these events over years and sites will illustrate individual models' characteristic behaviour in seasonal transitions, and their strength regarding process interactions, in combination with ordinary multivariate analysis techniques.

Finally, sensitivity tests for the model parameters are planned to quantify the effect of parameter sensitivity on the models' reproducibility. "

- New Figure 9

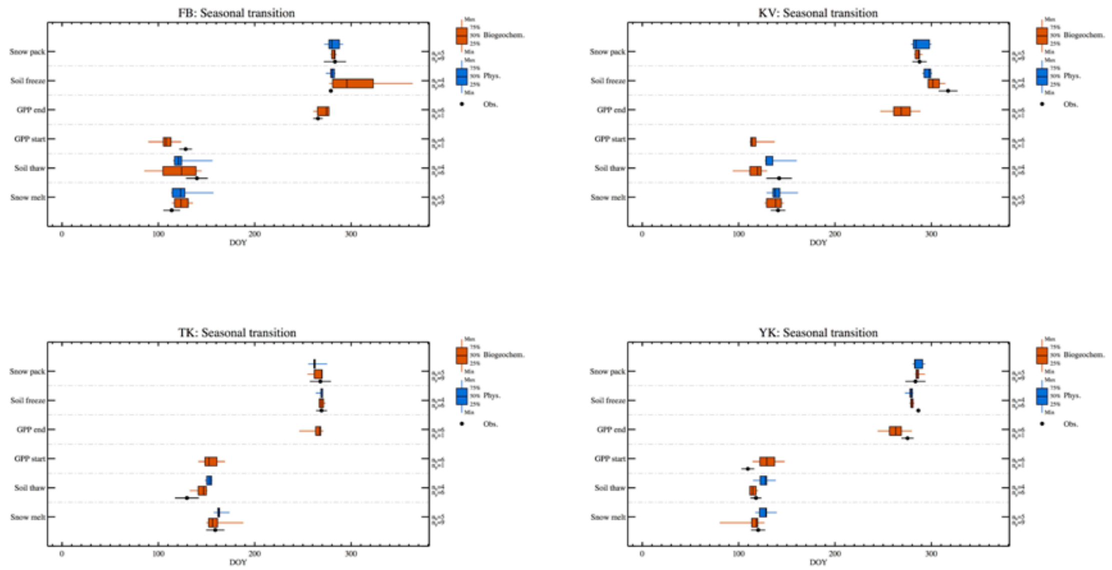

Figure 9. Example of seasonal transitions in ground temperature, snow, and vegetation among models.