Final thesis report Date : 6th of April 2016

Small-scale orographic gravity wave drag in stable boundary layers and its impacts on synoptic systems and near surface meteorology

A. Tsiringakis and G.J. Steeneveld and A.A.M Holtslag
Wageningen University, Wageningen, Netherlands

At present atmospheric models for weather and climate use enhanced turbulent mixing formulations under stable conditions, because the observationally based (short-tail) mixing function lacks the necessary momentum drag to accurately represent cyclonic filling over land. This enhanced mixing function (also known as the long-tail), introduces momentum drag that cannot be physically justified and deteriorates the score for near surface temperature, wind and boundary-layer height, and as such affects the forecast of phenomena like fog and frost. Here, orographic gravity wave drag for stable boundary layers is hypothesized to provide the missing drag needed without the disadvantages of using an enhanced mixing function. We include an updated parametrization in the WRF model (version 3.5) that represents the gravity wave drag induced by small-scale orography within the stable boundary layer, and adds it to the turbulent drag induced by a short tail mixing function.The schemes were evaluated for sixteen 8-day forecasts over the Atlantic and Europe in a winter situation. We conclude that the updated parametrization is able to reproduce sea level pressure, 10m wind and the cyclonic core pressure with higher accuracy than the other two setups. Cyclonic core pressure bias is reduced by 40% to 80% compared to the short-tail setup, and sea level pressure bias is reduced by up to 1 hPa (30%) over the whole domain, resulting in even smaller biases than the long-tail scheme. These results confirm our hypothesis that small-scale gravity wave drag may explain the need for a long tail function. Near surface wind bias is reduced by up to 40% compared to the long-tail and up to 20% compared to the short-tail setup, while 2m temperature bias is slightly increased (10%).

Key Words: Orographic gravity wave drag,Stable boundary layer, WRF model

Received . . .

1. Introduction

The complexity of the phenomena and the feedback mechanisms of the atmosphere, oceans and land surface are impossible to be explicitly represented in a single model. Therefore modeling efforts, especially in numerical weather prediction(NWP), have been focused in predicting accurately the large-scale synoptic conditions. Efforts are directed towards a substantial representation of the atmospheric boundary layer (ABL), the part of the atmosphere that is closest to the surface and interact with both land and ocean. These efforts have advance our understanding of the daytime ABL. When we have relatively calm conditions with low winds, which are typical occurring during winter nights, turbulence is significantly reduced and the boundary layer can become stratified (Holtslag *et al.* 2013; Steeneveld 2014). During these conditions we can identified the boundary layer as a Stable Boundary Layer (SBL). Recent studies have been very essential in understanding the mechanisms that govern the spatial and temporal variability of the SBL. The accurate representation of SBLs in NWP models has also related consequences to applications. Many phenomena that occur within the SBL (night frost, fog, low level jets) can affect wind energy production, aviation transportation sector and other sectors relevant to society that impact quality of life.

Despite many efforts most NWP models are unable to accurately forecast the SBL. The main problem originates from the parametrization of sub-grid scale processes that are not always well understood (Walsh *et al.* 2008; Kleczek *et al.* 2014). The interdependencies between these processes and their complexity inhibit the development of a robust and widely accepted parameterization for the SBL (Beljaars and Holtslag 1991; Beljaars and Viterbo 1998). Also Holtslag *et al.* (2013) found that most NWP models have difficulties distinguishing between very stable and weakly stable flows and consequently different SBL archetypes.

The model representation and the accuracy of the key variables in the SBL are therefore strongly influenced by the amount of turbulent diffusion that is imposed by the mixing length function (Cuxart *et al.* 2006; Sandu *et al.* 2013). So far there is not a universally accepted function for parameterizing turbulent diffusion and therefore the amount of turbulent mixing that is introduced within the SBL. The mixing functions that are

currently used in NWP model could be classified into two categories: a) the short-tail formulation and b) the long tail formulation.

The short tail formulation, which is based on Monin-Obukhov similarity theory introduces near zero mixing under stable conditions (Richardson number > 0.2). This formulation has been verified by observational and modeling studies with large eddy simulation models (Beljaars and Holtslag 1991; Mauritsen and Svensson 2007; Svensson and Holtslag 2009). Models applying this short tail formulation usually lack the sufficient cyclonic filling to accurately represent the development and life cycle of cyclones above land (Louis 1979; Holtslag 2006). Beljaars and Viterbo (1998) showed that under very stable conditions short-tail schemes also have significant near temperature bias that occur due to radiative runway cooling at the surface. The runaway cooling is a result of the unrealistic decoupling of SBL from the surface layer due to insufficient mixing (Sandu *et al.* 2013).

To improve these errors, Louis (1979) introduced an enhanced mixing function for the mixing length, which is called a long-tail formulation. This function was revised by Viterbo *et al.* (1999) and is currently used by the ECMWF forecast model. The long tail mixing function is known to increase the exchange coefficient for momentum and heat under stable conditions, resulting in enhanced turbulent drag that strengthens the Ekman Pumping Mechanism (EPM). This results in more cyclonic filling and better score for cyclonic systems core pressure (CCP), but distorts the structure of temperature and wind profiles in the ABL because it introduces more turbulence than is justified (Steeneveld *et al.* 2008; Sandu *et al.* 2013). This problem accounts for the biases in the 2m temperature, 10m wind and the location of the low-level jet and boundary-layer height over northern latitudes, especially during winter, and are present in many NWP model and Global Climate Models (GCM) (Cuxart *et al.* 2006; Walsh *et al.* 2008; Steeneveld *et al.* 2008; Sandu *et al.* 2013; Holtslag *et al.* 2013).

Orographic gravity wave drag (OGWD) for unresolved sub-grid scale orography, was introduced as a theory by Lilly (1972) and first parametrized by Palmer *et al.* (1986), who showed that including OGWD in the NWP models resulted in increased accuracy of synoptic wind (westerlies) and sea level pressure. Since then the theory has been further developed by Kim and Arakawa (1995) and Kim and Doyle (2005), and it has been implemented in the Weather and Research Forecasting (WRF) model. Many studies (Kim and Doyle 2005; Hong *et al.* 2008; Sandu *et al.* 2013; Choi and Hong 2015) have shown that including the OGWD parametrization does improve the accuracy of the models on the scale of large-mountain effects on the momentum budget of the troposphere, but its does not fully account for the missing drag within the SBL.

In order bridge this gap, Steeneveld *et al.* (2008) hypothesized that the missing drag can be generated from small-scale orography within relatively shallow SBLs. This hypothesis, developed through many studies (Chimonas and Nappo 1989; Nappo 2002; McCabe and Brown 2006), inspired the study of Steeneveld *et al.* (2008), who combined this new parametrization (GWDSBL) with the original short-tail formulation and managed to show that the drag generated by this setup was close to the results obtained from the artificially long-tail formulation without deteriorating the accuracy of the near surface wind, the LLJ and the boundary-layer height. The findings of Steeneveld *et al.* (2008) have also been confirmed by the studies of Sandu *et al.* (2013) and Lapworth (2015); Lapworth *et al.* (2015).

The updated scheme of Steeneveld *et al.* (2008) was tested in a single-column model, but its accuracy has not been tested yet in an fully 3D operational model for an extended time period. In this study, we implement this scheme in the WRF model (version 3.5) Skamarock *et al.* (2008) and test it against the short tail and long tail formulation. The main research goal is to investigate if the new scheme (GWDSBL) can effectively function as a long-tail scheme (LT) in terms of cyclonic filling for synoptic systems and 2m temperature score, while it retains the accuracy of a short-tail scheme for near surface meteorological variables. Section 2 presents some theoretical background on mixing functions and gravity wave theory, while methodology is presented in section 3. Experimental Results and Discussion appears in section 4 and 5 and Conclusions are drawn in section 6.

2. Theoretical Background

2.1. *Turbulence parameterization in the SBL*

In order understand the influence of the OGWD in the momentum budget we rely on a prior knowledge of the difference between the short-tail and long-tail schemes in terms of stability function (ϕ_m) and the diffusion coefficient (K_m) for momentum for in the SBL. In this study we will utilize a K-profile method to estimate the diffusivity. The equation for the diffusion coefficient is given by Troen and Mahrt (1986) and Holtslag and Boville (1993) and it reads:

$$K_m = \frac{k u_* z}{\phi_m} \left(1 - \frac{z}{h}\right)^2 \tag{1}$$

Here k is the Von-Karman constant ($k = 0.4$), z is the model level height, h is ABL height, u_* is the friction velocity in $m\ s^{-1}$ and ϕ_m is the stability function. The stability function under a stable regime reads (Dyer 1974; Stensrud 2007):

$$\phi_m = 1 + a\frac{z}{L} \tag{2}$$

Herein L is the Obukhov length and a is stability parameter with the value 5 for a short-tail scheme and 2 is used as a proxy for the long-tail scheme. These functions are included in the YSU boundary-layer scheme, which we will utilized in the short-tail and long-tail setups.

2.2. *Orographic gravity wave drag parameterization for the SBL*

To calculate the drag induced by the GWDSBL scheme (τ_{wave}) we utilized the following non-linear equation introduced by Nappo (2002) and used by Steeneveld *et al.* (2008). The equations reads:

$$\tau_{wave} = \begin{cases} \frac{1}{2}\rho_0 k_s (UH)^2 \sqrt{\frac{N^2}{U^2} - k_s^2}, & \text{if } \frac{N}{U} \geq k_s. \\ 0 & \text{if } \frac{N}{U} < k_s. \end{cases} \tag{3}$$

Under the assumption of weak wind, the equation is reduced to a linear equation Nappo (2002) :

$$\tau_{wave} = \frac{1}{2}\rho_0 k_s H^2 N U \tag{4}$$

In equation 4 ρ_0 is the air density in $kg\ m^{-3}$, U is the mean background wind at the top of the PBL that is perpendicular to the orography, H is the orographic amplitude of the subgrid-scale orography, N is the Brunt-Vaisala frequency and k_s is the orographic wave number. The study of Steeneveld *et al.* (2008) and our sensitivity tests show that the difference between the non-linear and the linear version for the gravity wave drag have minimal differences in the final results (not shown). Therefore, we

decided to use the linear approximation in the parametrization of the gravity wave drag within the SBL, due to its lower complexity. To account for the divergence of gravity wave drag as we approach the top of the SBL and its reduction to near zero values at the top of the SBL, we used the following equation (Steeneveld *et al.* 2008):

$$\tau_{wave}(z) = \tau_{wave}(0)\left(1 - \frac{z}{h}\right)^2 \quad (5)$$

This shape was inspired by a high resolution timescale model simulation over hills.

The updated parameterization takes into account the effect of gravity wave breaking and overturning due to wave saturation only in the momentum budget of the horizontal wind and does not account for perturbations in temperature, pressure or vertical wind that are caused by gravity waves(Nappo 2002). We have implemented this parameterization as an extension on the original OGWD parameterization of Kim and Doyle (2005) for the WRF-ARW model.

2.3. *Subgrid-scale orographic amplitude*

The amplitude of the subgrid-scale orography and the orographic wave number (k_s) are necessary for the calculation of gravity wave drag in equation 4. In the original study of Steeneveld *et al.* (2008) with the single-column models, H and k_s were constant in space. In order to make the parametrization more realistic we decided to implement a calculation of the subgrid-scale orography amplitude and the orographic wave number with the algorithms of Kim and Doyle (2005). This means that H and k_s are calculated by the following equations Kim and Doyle (2005):

$$H = 2\sigma_h \quad (6)$$

$$k_s = \frac{(1 + L_x)^{(1+OA)}}{\lambda_{eff}}, \quad (7)$$

With σ_h being the subgrid-scale orographic standard deviation and OA is the orographic asymmetry (non-dimensional) that depends on the asymmetry of the subgrid-scale orography and its relative location to the model grid box. The effective orographic length L_x (non-dimensional) is the subgrid-scale mountains width integrated over the grid cell and normalized by the gridcell size and λ_{eff} is the effective grid length (in m).

H and k_s are calculated in every gridcell of the domain and the usually take values of 0-100m for H and $4 * 10^{-5}$ to $8 * 10^{-5}$ m^{-1} for k_s, but H can have values up to 1000m in areas with strong heterogeneity. To ensure that we do not include larger values of H in the calculation of τ_{wave} we restricted the parametrization to be active only in gridcells, where the ABL height is larger than $2\sigma_h$. With this restriction we ensure that only small-scale orographic features are used to calculate τ_{wave}, resulting in realistic values of gravity wave drag in the SBL and an increase in model stability.

3. Methodology

3.1. *Model description and experimental setup*

In our study we have used the WRF model Version 3.5.1. WRF is a mesoscale meteorological model and we will use it together with the dynamic core option of the model ARW (Skamarock *et al.* 2008). We use a horizontal resolution of 25x25 km in a domain with 315x226 gridcells (Figure 1) and a Lambert projection. The vertical resolution for our model is 34 levels with a model top at 100 hPa and 15 levels is the below 900 hPa to promote the representation of the ABL. This configuration allows us to simulate the effects of gravity wave drag effectively and also improve the detection of LLJ events in the SBL. Each model run is initialized at 12UTC and lasts for 8 days (192 hours) with an adaptive model time step. Initial conditions by ECMWF operational analysis data at a 0.25^ox0.25^o resolution are provided at the start of each run and as boundary conditions every 6 hours.

The whole experiment consists of 16 model runs in the winter of 2011/2012, with the first run initialized in 15 December 2011 at 1200 UTC and the last one initialized at 14 January 1200 UTC. This results in consecutive 8-day runs that are initialized every 2 days and cover the period of more than a month. We choose to use this specific period due to a significant amount of cyclonic and anticyclonic systems crossing/(developing above) Europe. Also, the cold conditions during these winter months will be ideal for the creation of SBLs that are necessary to produce sufficient amounts of gravity wave drag.

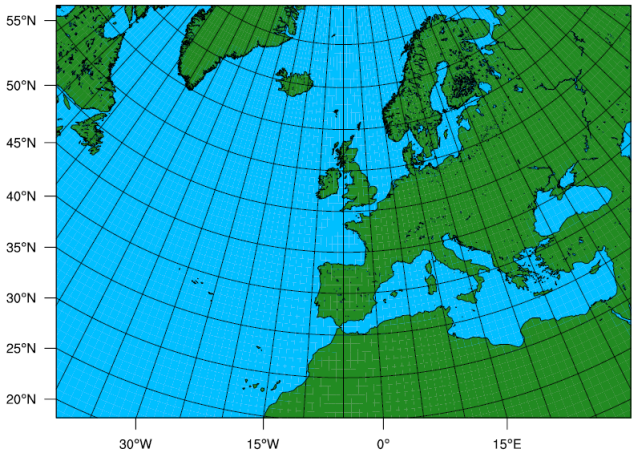


Figure 1. The domain used in WRF model runs. It consists of 315 by 225 gridcells with a 25 km horizontal resolution and 34 vertical levels.

To account for the orographic gravity wave drag in stable boundary layer we will use the parametrization that was introduced in section 2. Our experiment consist of 3 model setups: a) short tail mixing function, b) short tail mixing function with the new GWDSBL parametrization included and c) long tail mixing function. All 3 model setups include the OGWD option of the dynamical core in the WRF-ARW model. Our goal is to compare these 3 setups and check if the new GWDSBL setup can reproduce the synoptic conditions (SLP and cyclonic life cycle) with accuracy similar to the long-tail (LT) setup, while not reducing the accuracy in 10m wind and 2m temperature compared to the short-tail (ST) setup.

We have a set of default settings regarding the parametrization schemes of WRF. To represent the boundary layer we used the Yonsei University (YSU) scheme Hong *et al.* (2006), which is a commonly used and well tested scheme (Gilliam and Pleim 2006). YSU scheme can be used in combination with the OGWD option of the dynamical core in WRF model Shin *et al.* (2010). For microphysics scheme we used the WSM-3 simple ice scheme (Hong *et al.* 2004). For surface layer parametrization we used the MM5 Monin-Obukhov scheme (Jimenez *et al.* 2012, a), for the land surface the unified Noah land-surface model Chen and Dudhia (2001) and the Kain-Fritsch scheme was used for cumulus parameterization (Skamarock *et al.* 2008). For long wave and shortwave radiation we have used the RRTM scheme (Mlawer *et al.* 1997) and the Duhdia shortwave radiation scheme (Dudhia 1989) respectively.

We have conducted a series of sensitivity tests in two model runs. The first case was initialized the 15 of December and the

31 of December. The sensitivity tests that we did include a series of changes in the parameterization (linear vs non-linear equations for gravity wave drag), shortwave radiation scheme, initialization time and amount of vertical levels (34 and 61). These sensitivity test will not be analyzed in section 4 , but we will refer to them in the section 5 .

3.2. *Statistical analysis methods*

To evaluate the impact of the selected physical parameterization schemes, model results have been compared to the ECMWF operational analysis and Cabauw tower observations in the Netherlands. Sea level pressure, cyclonic core pressure, 2m temperature, 10m wind, relative humidity at 2m and the geopotential height of the 500 hPa field have been analyzed extensively and for all model runs, while other parameters like SBL height and LLJs has been addressed for specific case studies. The statistical analysis has been done with the use of conventional statistics from Willmott (1982). We have calculated the mean bias error (MBE), mean absolute error (MAE) and root mean squared error (RMSE) for each variable and we have check both the temporal and spatial evolution of errors for all three model setups. On top of that, we compared the CCP of certain cyclones with the ECMWF forecast in order to have a direct comparison with the integrated forecast system (IFS).

The SLP fields (lat-lon fields) showed systematic negative bias (-0.6 hPa averaged over the whole domain) in the initial time step, which can result from differences in the orographic height. These errors can also occur from differences in the calculation of SLP between ECMWF(explicit) and WRF (implicit) or due to problem when calculating SLP above snow covered mountains in the WRF model. We also found systematic error (30 m averaged over the whole domain) in the calculation of HGT_500 hPa (lat-lon fields). The error source could be due to orographic difference between WRF and ECMWF or due to the interpolation that we did to identify the height of the 500 hPa field from the WRF output, because WRF has no constant pressure fields. Since we are not interested in systematic errors in the model setups, we have subtracted the errors, obtained from the differences of WRF and ECMWF at time step 0, from each of the following time steps for both variables.

Understanding and measuring the performance of 3 separate model setups and for 5 different variables can be very difficult with separate figures for each variable. Therefore, we decide to show the measure of accuracy for each variable and for all model setups with the use of the Taylor diagram Taylor (2001). From the Taylor diagram we do a summarizing analysis for all 16 model runs with the use of the Taylor diagram model score S which reads (Taylor 2001) :

$$S = \frac{4(1+R)^4}{(\sigma + \frac{1}{\sigma})^2(1+R_o)^4} \tag{8}$$

Where R is the correlation coefficient calculated between the WRF setups and ECMWF operational analysis, σ is ratio of the RMS difference between WRF setups and ECMWF and R_o is the limiting correlation taken as 1 for all model setups. S takes values between 0 and 1 with 1 being a perfect score.

4. Experimental Results

From the 16 different case studies we selected for detailed analysis the model run of 2-10 of January 2012. After the sensitivity analysis we concluded that this case is suitable to show the effects of the new GWDSBL parametrization in all of the variables. The parameters that are included in the analysis that follows are the cyclonic core pressure(CCP), sea level pressure(SLP), 2m temperature (T2m) and 10m wind(U10m). Model scores for geopotential height at the 500 hPa level (HGT_500 hPa) and relative humidity at 2m (RH_2m) will be addressed.

4.1. *Evaluation of Cyclonic Core Pressure*

The first meteorological parameter we investigate is the cyclonic core pressure (CCP). The impact of boundary layer stress on the cyclone life cycle and CCP has been confirmed by many studies (Anthes and Keyser 1979; Beare 2007; Plant and Belcher 2007; Boutle *et al.* 2009, 2015) and is hypothesized to be connected to the EPM. The EMP is responsible for the cyclonic filling due to vertical wind and ventilation on the ABL top (Holton 1992; Beare 2007). Beljaars and Viterbo (1998) showed that in order to have good score for cyclonic filling an enhanced mixing function was necessary. Contrary to their results we show that the new GWDSBL parametrization is able to accurately reproduce the CCP of cyclones due to enhanced momentum drag from gravity wave dissipation within the SBL.

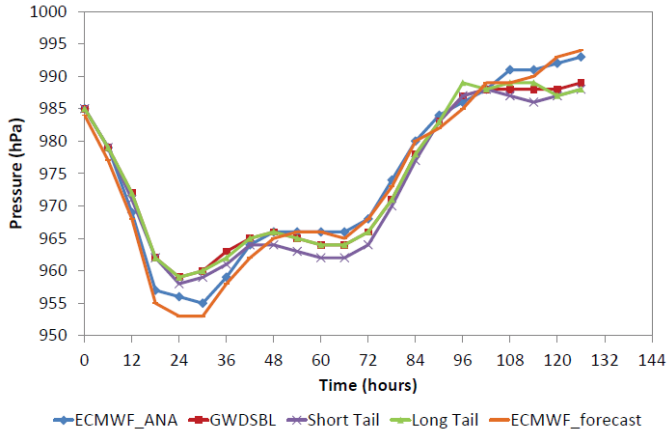
For this reason we analyzed of the life cycle of two separate cyclonic systems (Figure 2). System 1 (Figure 2a) is a cyclone that develops above the North Atlantic region and its trajectory (Figure 2b) passes above England and Scandinavia. The second system (Figures 2c, 2d) forms above the Dalmatian coast, after 84 hours in the simulation, and has a short life cycle of 48 hours before it decays above the Aegean sea.

For cyclone 1 (Figure 2a) all three WRF experimental setups and the ECMWF forecast can represent the life cycle of the cyclone with very good accuracy (Table 1), taking into account that model runs differentiate substantially (3-5 hPa) from the observed CCP between 12 and 36 hours. GWDSBL has been able to reproduce the cyclonic core pressure with higher accuracy than the LT and ST setups. In Table 1 one can see that the GWDSBL outperforms ST in every statistical score (MBE,MAE and RMSE) and shows similar results with the LT setup. Compared to the ST experiment we found an improvement of 0.5 hPa in MBE and up to 1 hPa for MAE and RMSE. These results support our hypothesis that the new GWDSBL parametrization functions as a long-tail mixing function. ECMWF forecast, which also uses a long-tail mixing function, has the higher accuracy for system 1 compare to the rest of the model simulations. This was expected since the ECMWF model is tuned specifically for accurate medium range forecasts and operates with significant higher vertical model levels that enhance the accuracy of the model.

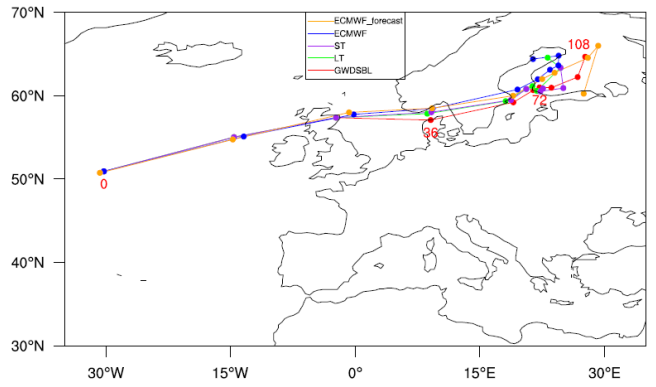
Table 1. Statistical errors of cyclonic core pressure for all our WRF setups and ECMWF forecasts for cyclone 1 and 2 for the simulation of 2-10 January 2012. Comparison has been done against ECMWF operational analysis data.

Cyclone 1	MAE	MBE	RMSE
GWDSBL	2.23	-0.23	2.72
ST	2.72	-1.27	3.25
LT	2.27	-0.18	2.79
ECMWF forecast	1.13	-0.86	1.40
Cyclone 2	MAE	MBE	RMSE
GWDSBL	3.22	-3.00	3.96
ST	6.11	-5.89	7.64
LT	6.00	-6.00	7.63
ECMWF forecast	4.89	0.22	6.03

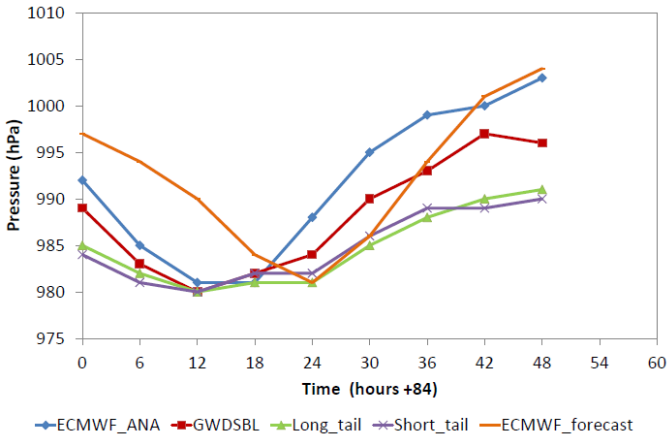
Figure 2c shows the development of the CCP for system 2. GWDSBL setups can reproduce the track and the cyclonic life system better than LT,ST and ECMWF_forecast (Table 1). LT and ST setups show larger errors both for the CCP and for the track of the system. GWDSBL was also been able to outscore



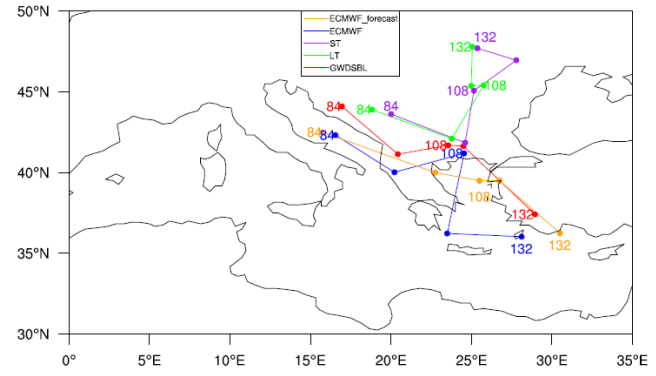
(a) Cyclone 1 CCP



(b) Cyclone 1 trajectories



(c) Cyclone 2 CCP



(d) Cyclone 2 trajectories

Figure 2. Modeled and observed cyclone core pressure (hPa) (Figures 2a and 2c) and trajectory (Figures 2b and 2d) for 2 cyclones in the simulation of 2-10 January 2012. We compare the 3 model setups (GWDSBL, ST and LT) with the operational analysis data of ECMWF and the ECMWF forecast data. Cyclone 1 is present from the start of the simulation , while cyclone 2 develops after 84 hours in the simulation. The location of each cyclone is identified every 12 hours (dots), the trajectory (lines) is done by connecting the locations of the cyclones and the numbers next to certain dots are showing the hours in the simulation.

ECMWF and represent the system with 1.5 hPa lower MAE and RMSE. For this cyclone it is clear that GWDSBL is more effective compared to the enhanced mixing function of the LT setup. On top of that, GWDSBL is able to reproduce the deepening phase of the cyclone with significantly higher accuracy compare to all other setups, an indication that the new parametrization could improve the accuracy of the forecasts for the whole life cycle of the cyclones. Regarding the trajectory of cyclone 2, the different tracks between GWDSBL and the other 2 model setups could have been caused also by the orography in the borders between Greece and Bulgaria. More specifically, cyclone 2 in the GWDBSL setup is located in the southern part of the mountain range, while in LT and ST setups the cyclone is located north of it. Therefore we hypothesize that the mountain range could have forced cyclone 2 to deviate north in ST and LT setups.

Our hypothesis for the improved cyclonic development of the GWDSBL compared to ST is based on the effects of boundary layer stress on the cyclonic life cycle. The enhanced momentum drag from the dissipation gravity wave within the SBL results in higher values of boundary layer stress compared to a short tail mixing function. As a result of the enhanced stress in the ABL, the near surface wind is decreased (see section 4.4), resulting in a strengthened EPM and faster spin down of the cyclone. From the analysis of the two systems, we can conclude that under strong stability ($Ri > 0.5$) and high values of subgrid-scale orographic amplitude the momentum drag induced by the GWDSBL setup can be stronger than that of long-tail mixing function (LT) (Figure 2c). Consequently the cyclonic filling rate is faster and we can see a better representation of the CCP for the GWDSBL setup. The behavior of the GWDSBL parametrization has been also tested for three other cases with cyclonic activity above Europe (not shown), which verified the superiority of the new parametrization compared to the ST one and its similarity to the LT setup in terms of CCP and cyclonic filling rates.

4.2. *Sea level pressure*

Next variable in our analysis is the domain-wide sea level pressure. SLP is strongly affected by the accuracy in the synoptic meteorological conditions. Therefore we expect to see the impact of the representation of the cyclonic life cycles (see section 4.1).

For the GWDSBL experiment (see Figure 3a) we found significant negative bias in the order of 4 to 5 hPa above Europe, which is increased by 1 hPa above region with complex orography. Positive bias can be observed in the North Atlantic region with a magnitude of 2 hPa and above Greenland (8-10 hPa). The negative bias above Europe is an effect of a weaker high pressure ridge that is build between the Azores high pressure system and the Siberian high during the second half of the simulation. In ECMWF operational analysis data this ridge results in significant high pressure across Italy, the Balkan region and east Russia. In the WRF simulation this ridge is substantially weaker (up to 10 hPa), which results in more cyclonic activity over Greece and deeper low pressure systems (figure 2c). Stronger than observed high pressure west of Portugal and a weaker Iceland low is the reason for the positive bias in the north Atlantic. The positive error in small region above Greenland can be due to errors in the

calculation of systematic error from time step 1, but we are certain that this doe not influence the model setups comparison.

Figures 3b and 3c show the spatial distribution of the differences in MBE and RMSE, averaged over the whole period (2-10 Jan), between the GWDSBL, LT and ST setups. In these figures we have subtracted the MBE of LT and ST from the MBE of GWDSBL setup. By doing so we can clarify whether the new scheme is an improvement compared to the LT and ST schemes. Figure 3b shows this spatial difference of the two errors between GWDSBL and LT. It is clear that the new GWDSBL parametrization reduces the MBE and RMSE by 1 to 4 hPa (20%-30%) in central and eastern Europe due to the more accurate representation of the high pressure ridge and cyclone 2. A small deterioration of SLP accuracy, up to 1 hPa, is visible in northern Africa and around Iceland. We did a similar comparison between GWDSBL and ST experiments (Figure 3c). In this case the improvement in the accuracy of SLP above Europe increases by 3 to 5 hPa (30% to 50%), while the deterioration above northern Africa and Iceland increases to 1 hPa.

In addition to the analysis of the spatial evaluation of MBE and RMSE mean errors we evaluated the temporal development of the model scores averaging the errors over the whole domain with a 6 hour interval (Figure 4). In the first half of the simulation (until 84 hours) the statistical values for all three model setups show small differences. MBE shows small positive bias (0.5 hPa), while MAE and RMSE are increasing. During the second half of the simulation the model setup scores for SLP start to deviation from each other. The MBE for GWDSBL reaches near zero values with differences of 1 to 2 hPa over the whole domain, compared to LT and ST. Smaller MAE and RMSE for GWDSBL are also visible in the same Figure 4.

The difference in the accuracy of the SLP fields and the reduction of the statistical errors above Europe and Russia in the GWDSBL run are due to a better representation of individual low pressure systems trajectory and intensity between the GWDSBL and the other two setups. This improvement can be attributed to the enhanced momentum drag from gravity wave dissipation that is represented in the GWDSBL parametrization. In GWDSBL setup the CCP and the track of the systems is significantly more accurate than that of LT and ST, in which cyclone 2 deviates north and cross Eastern Europe, resulting in a distortion of the ridge between the two high pressure system and consequently a reduction of SLP accuracy compared to GWDSBL and ECMWF.

4.3. *Evaluation of 2m temperature*

Next we evaluate the scores for 2m temperature. Near surface temperature in the SBL depends on the amount of diffusion that we represent in each of the model setups. Beljaars and Viterbo (1998) showed that the enhanced mixing functions, similar to the mixing function of our LT setup, have higher near-surface temperatures compared to traditional "short-tail" mixing functions due to enhanced mixing. This results in smaller negative bias in 2m temperature for "long-tail" mixing functions. Taking that into account, we expect that the LT setup will show smaller biases than ST and GWDSBL setups.

We will start our analysis with Figure 5a, which shows the spatial distribution of MBE and RMSE for GWDSBL. The MBE for temperature above Europe is -2 to -3 K compared to ECMWF operational analysis. Statistical errors increase up to -10 K above regions with very stables conditions with the highest MBE appearing in northern Finland and Russia. In the Sahara region all model setups show negative bias in night temperatures and positive bias in daytime temperatures. The negative biases of the WRF model in northern latitude has been reported also from Cassano *et al.* (2011), where they found strong

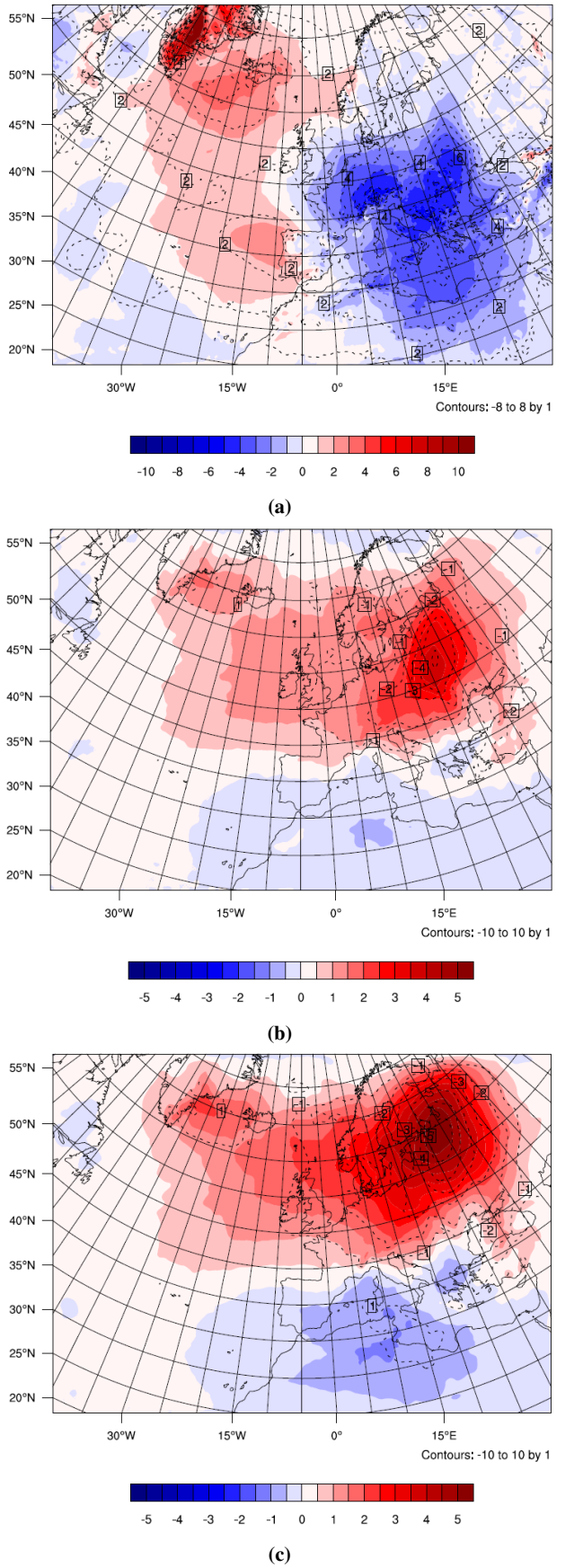


Figure 3. Spatial distribution of mean statistical errors of SLP compared to the ECMWF operational analysis data in the simulation period of 2-10 January 2012. Figure 3a shows the MBE (filled contours) and the RMSE (dotted contours) of SLP for the GWDSBL setup. Figure 3b shows the differences of MBE (filled contours) and of RMSE (dotted) between the GWDSBL and LT (GWDSBL-LT). Figure 3b shows the differences of MBE (filled contours) and of RMSE (dotted) between the GWDSBL and ST (GWDSBL-LT). For figures 3b and 3c positive values of the difference in RMSE shows improvement with the GWDSBL setup. For the difference of MBE positive values show improvement with the GWDSBL if the MBE of GWDSBL (3a) setup is negative and deterioration if the MBE of GWDSBL is positive.

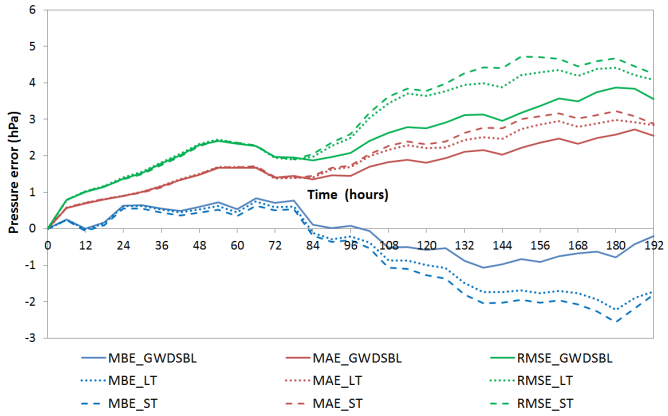


Figure 4. Temporal evolution of the mean statistical errors for sea level pressure (averaged over all gridcells) for each of the model setups (GWDSBL,ST and LT) in the simulation 2-10 of January 2012.

negative bias between -0.4 to -4.8 K for WRF model over a pan-Arctic simulation. Many studies (Beljaars and Viterbo 1998; Steeneveld *et al.* 2008; Holtslag *et al.* 2013; Sandu *et al.* 2013) have confirmed that models using "short-tail" functions are more susceptible into creating a decoupling of the surface layer from the rest of boundary layer under stable conditions. The decoupling strengthens the radiative cooling at the surface, which results in strong biases in near surface temperatures.

Figure 5b shows the difference of MBE and RMSE between GWDSBL and LT. One can see that GWDSBL shows stronger bias in 2m temperature by up to -2 K depending on the prevailing stability conditions in each region. Above Europe the difference is in the order of -0.5 K and rises up to -1 K in mountainous regions. Figure 6 shows that the differences between GWDSBL and LT are visible from the first time steps and are gradually increasing towards the end of the simulation.

The difference between GWDSBL and ST is also visible in figure 5c. The difference in MBE and RMSE is in the order of -0.5 K above Europe and up to -2 K in regions where we have prevailing stable conditions in the boundary layers (Greenland, Russia). The temporal evolution of the model scores (Figure 6) shows that the difference in the 2m temperature bias between GWDSBL and ST do not appear until near the end of the simulation time and the total difference over the whole domain is in the order of 0.5 K. This reduction in accuracy was not found in the study of Steeneveld *et al.* (2008) but was found by Choi and Hong (2015). The later authors attributed this to lower more accurate wind speeds (see section 4.4) in model runs that resulted in less mechanical turbulence and small surface fluxes, when they included the gravity wave drag parametrization.

Table 2. Statistical errors of 2m temperature (K) for all our WRF setups at the location of Cabauw tower. Comparison has been done against the observational data from Cabauw tower.

Experiment	MAE	MBE	RMSE
GWDSBL	1.91	-1.69	2.64
ST	1.57	-1.36	2.35
LT	1.73	-1.59	2.25

Our finding for the spatial distribution of errors also agree when we zoom in specific locations. For the 2m temperature at Cabauw station in the Netherlands (Figure 2), we found that GWDSBL experiment has the lowest accuracy compared to the observations, but has only very small deviation from the ST setup. The difference between the observation and the model setups originate mainly from one time period (the night of 8-9 January) with very low wind speeds around 2 ms^{-1} at the height of 10m,

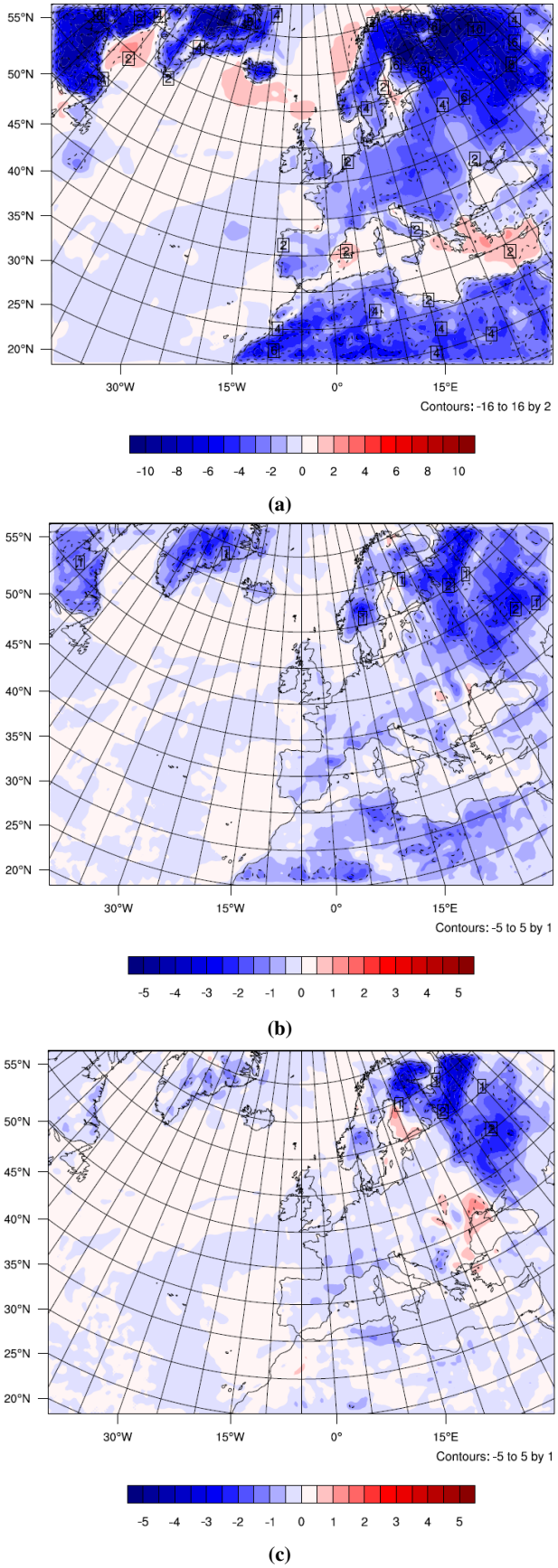


Figure 5. Spatial distribution of mean statistical errors of 2m temperature compared to the ECMWF operational analysis data in the simulation period of 2-10 January 2012. Figure 3a shows the MBE (filled contours) and the RMSE (dotted contours) of T2m for the GWDSBL setup. Figure 3b shows the differences of MBE (filled contours) and of RMSE (dotted contours) between the GWDSBL and LT (GWDSBL-LT). Figure 3b shows the differences of MBE (filled contours) and of RSME (dotted contours) between the GWDSBL and ST (GWDSBL-ST). For figures 3b and 3c positive values of the difference in RMSE shows improvment with the GWDSBL setup. For the difference of MBE positive values show improvement with the GWDSBL if the MBE of GWDSBL (3a) setup is negative and deterioration if the MBE of GWDSBL is positive.

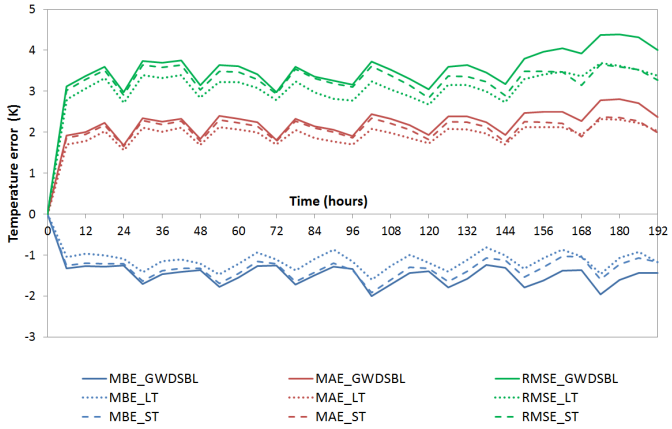


Figure 6. Temporal evolution of the mean statistical errors for 2m temperature (averaged over all gridcells) for each of the model setups(GWDSBL, ST and LT).

which resulted in very stable conditions and temperatures below the freezing point for all model setups. During the same time the observations show similar values for 10m wind speeds, but 4 K higher 2m temperatures. The fact that 10 wind speed is similar for both model results and observations, and the T2m values do not differ between the model setups, lead us to hypothesis that the deviations of 2m temperatures are caused by enhanced radiative cooling in the models setups for that specific night. The source of the enhanced radiative cooling is likely to be the absence of clouds in the model runs during that night, while in the observations we have low level clouds and small amounts of precipitation that reduces the net radiation budget at the surface.

The deviation in the 2m temperature scores above Russia can be attributed to two different sources. The first one is the difference in the mechanical turbulence that are caused by the lower wind speed in GWDSBL experiment in that region (see section 4.4, figures 7b and7c) or due to different synoptic meteorological conditions (see section 4.2). Regarding the first source we found that the friction velocity (u_*) for GWDSBL is significantly lower compared to ST and LT in these regions. These is a result of the strong deviation between the two setups due to the fact that GWDSBL run has lower wind speed compared to the other two setups. Having a look at equations 1 and 2 we will see that the amount of the diffusivity is proportional to the value of u_* and also the stability function since it includes the Obukhov length, which is related to the friction velocity. Therefore the regions with lower friction velocities tend to have smaller values of K_m and as a result weaker mixing near the surface. In reality the deterioration of the results that occur due to this mechanism is not only drawback of the GWDSBL. Section 4.4 shows that the lower near surface wind speeds for GWDSBL experiment are actually closer to the observations compared to ST and LT experiment. This indicates that LT and ST experiments show compensating errors in 2m temperature by having unrealistic amounts of friction velocity near the surface that results in enhanced wind-induced mixing.

The difference in 2m temperature could have been caused also by the differences in the synoptic conditions that we described in 4.2 and 4.1. In this section we showed that the synoptic conditions in north eastern Europe are different for LT and ST experiment, compared to GWDSBL experiment, due to fact that cyclone 2 is moving north towards Russia instead of remain above Greece. This cyclone could have caused weaker 10m wind speed for GWDSBL resulting in weaker mixing in the SBL as we explained before. Also advection of warm air from the Balkans or more cloud cover over Russia for LT and ST setups could have been caused by this cyclone.

4.4. *Evaluation of 10m wind*

The next parameter that we evaluate is 10m wind speed. Figure 7a shows the spatial distribution of the statistical errors for GWDSBL run compared to the ECMWF operational analysis data. We find that GWDSBL overestimates 10m wind speed in the order of 2 to 4 at regions with strong orographic complexity. Smaller biases around 1 ms^{-1} are visible in regions with small-scale orography. Due to the westerly direction of the prevailing wind we can see that overestimation of the near surface wind is also located in the location east of mountain ranges (East Mediterranean Sea, west coast of Greenland). A negative bias is only visible in some parts of the Atlantic Ocean, the Sahara desert and Russia. The extent of the negative bias is generally smaller and the prevailing bias for the 10m wind speed over the whole domain is positive (8). The statistical errors show a gradual increase over time after the rapid increase that occurs within the first 6 hours (Figure 8).

If we compare the model scores for 10m wind speed of GWDSBL with the ones of LT scheme (Figure 5b) we find a clear improvement in the 10m wind field when we use the GWDSBL setup. The MBE over the whole domain (Figure 8) is reduced by 0.2 ms^{-1} to 0.4 ms^{-1} (20% to 40%) with the strongest reduction, around 2 ms^{-1}, occurring in regions with high values of the subgrid-scale orographic amplitude (H), because of the higher values for H increase τ_{wave} resulting in more momentum drag. The same comparison between GWDSBL and ST experiments shows also improvement with the use of the GWDSBL, in the order of 0.2 ms^{-1} to 0.3 ms^{-1}, for the whole domain and similar differences in MBE over mountainous regions.

The improvement in accuracy of 10m wind with the use of the GWDSBL can be attributed to 3 different sources. Between GWDSBL and ST setups the differences can occur either from impacts of the gravity wave drag parametrization, which enhances the momentum drag and therefore reduces the near surface wind speeds, or a better representation of the synoptic meteorological conditions with the use of the GWDSBL (section 4.2). In the comparison between GWDSBL and LT one should also take into account the difference in the estimation of the exchange coefficient for momentum as a result of using different parameter a in the stability function (Eq 2). The increase in the difference of the MBE and RMSE between GWDSBL and ST experiments after 72 hours in the simulation, implies that in the beginning of the simulation the difference in the mixing functions have an significant impact, while near the end of the simulation the deviations are mostly due to difference in synoptic conditions (see section 4.2) and the impact of the gravity wave drag parametrization in SBLs.

Table 3. Statistical errors of 10m wind speed for all our WRF setups at the location of Cabauw tower. Comparison has been done against the observational data from Cabauw tower.

Experiment	MAE	MBE	RSME
GWDSBL	1.21	0.29	1.59
ST	1.78	1.28	2.22
LT	1.37	0.64	1.75

We have also evaluated the near surface wind in the region of the Cabauw station with the use of observation data for all 3 model setups. Table 3 shows that the GWDSBL experiment has a 1 ms^{-1} smaller bias compared to the LT experiment and around 0.5 ms^{-1} improvement for MAE and RMSE. Compared to the ST setup the improvement in MBE is 0.4 ms^{-1} and around 0.2 ms^{-1} for MAE and MBE. These results are in agreement with the difference that we see in figures 7b and 7c over the Netherlands.

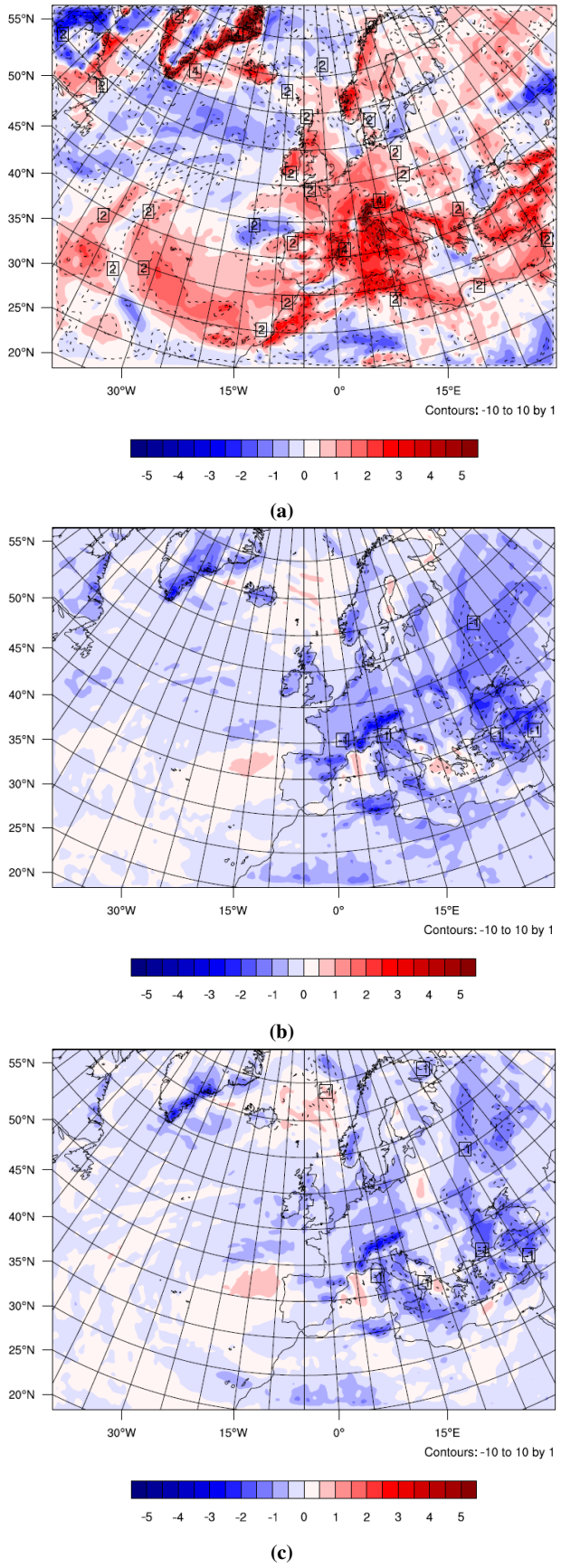


Figure 7. Spatial distribution of mean statistical errors of 10m wind speed ($U10$) compared to the ECMWF operational analysis data in the simulation period of 2-10 January 2012. Figure 3a shows the MBE (filled contours) and the RMSE (dotted contours) of U10 for the GWDSBL setup. Figure 3b shows the differences of MBE (filled contours) and of RMSE (dotted contours) between the GWDSBL and LT (GWDSBL-LT). Figure 3b shows the differences of MBE (filled contours) and of RMSE (dotted contours) between the GWDSBL and ST (GWDSBL-ST). For figures 3b and 3c positive values of the difference in RMSE shows improvement with the GWDSBL setup. For the difference of MBE negative values show improvement with the GWDSBL if the MBE of GWDSBL (3a) experiment is positive and deterioration if the MBE of GWDSBL is negative.

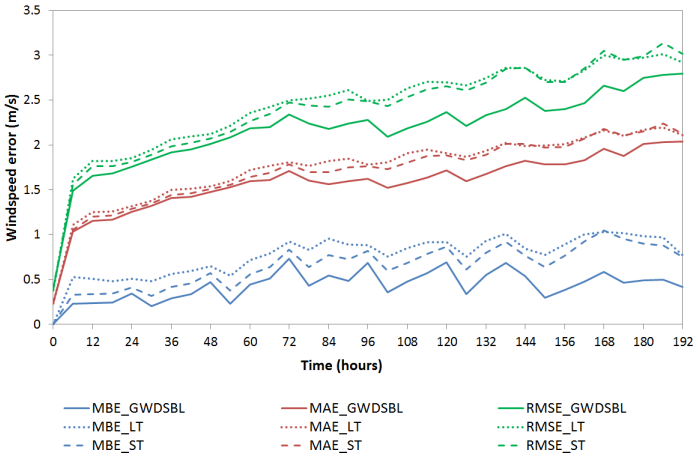


Figure 8. Temporal evolution of the mean statistical errors for 10m wind speed (averaged over all gridcells) for each of the model setups(GWDSBL,ST and LT).

4.5. Height of the 500 hPa pressure field

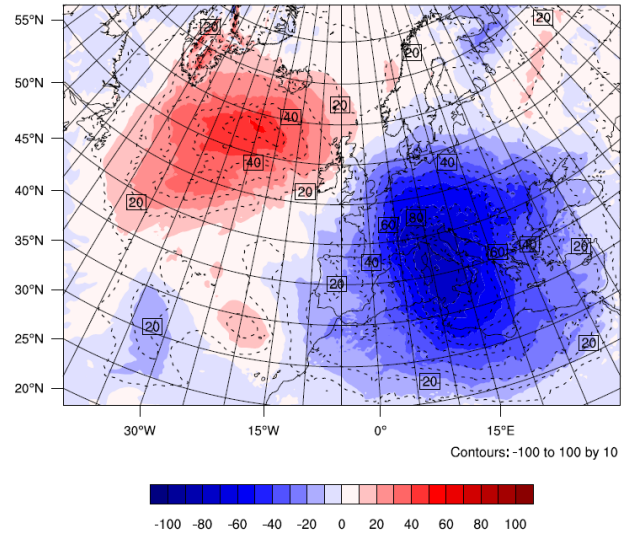
In this subsection we analyse the 500 hPa field score for the three model setup and we compare it with the ECMWF operational analysis values. The level of the 500 hPa field is considered one of the most important variables in evaluating the performance of an NWP model. Being closely connected with the accuracy in the representation of the barotropic circulation, HGT_500 hPa scores are used also by ECMWF to evaluate the time period under which a weather forecast has skill.

The spatial distribution of the statistical errors for GWDSBL setup is visible in Figure 9a. The region with the highest positive bias (40-50 meters) is located above the the Atlantic Ocean while a strong negative bias (60-80 meters) is visible above Southern Europe and Italy. By comparing the Figures 9a and 3a we found that the spatial distribution of positive and negative biases for SLP are also occurring in the HGT_500 hPa field. This indicates that the difference from the reference fields of the ECMWF operational analysis data is mainly due to deviations in the synoptic circulation, especially in the representation of the trajectory and intensity of cyclonic and anticyclonic system.

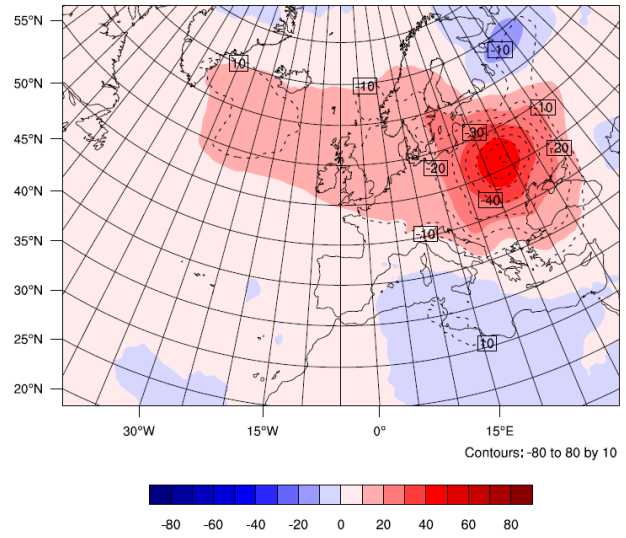
Figures 9b, 9c show similar pattern with Figures 3b and 7c for SLP. We identify that there is a reduction of the MBE up to 40 m (55 m for RMSE) above Easter Europe compared to the ST experiment, with an increase in errors above North Atlantic and North Africa in the order of 10 meters in MBE and RMSE. The improvement in the accuracy of the 500 hPa fields is due to better representation of the SLP and CCP in the GWDSBL setup compared to LT (see sections 4.1 and 4.2). Compared to the ST experiments GWDSBL shows a similar pattern with further reduction of in MBE up to 50m and RMSE above Europe and a slight in above North.

Regarding the temporal evolution of the errors (Figure 10) we find again a similarity with the corresponding SLP Figure 4. The statistical errors for all three model setup show no difference in the first 84 hours. On the contrary near the end of the simulation we can see that GWDSBL has -7 meters bias, which is 20 m smaller than that of LT and ST setups.

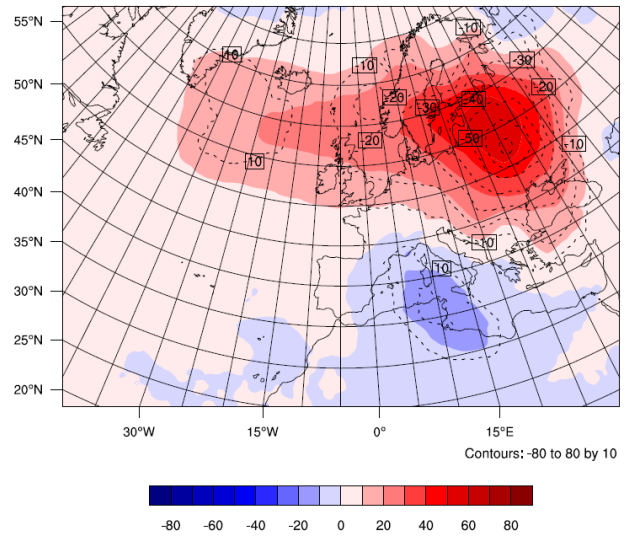
The difference in the HGT_500 hPa fields between the three model setups reveals that small modifications in the representation of the SBL can have larges impacts even at that such a height. This is quite striking if we take into account that the HGT_500 hPa variable is usually very robust. Similar finding have only been reported in the study of Sandu *et al.* (2013), where they have also found large differences in the accuracy of the HGT_500 hPa fields due to modifications in the mixing function in the SBL.



(a)



(b)



(c)

Figure 9. Spatial distribution of mean statistical errors in the height of the 500 hPa field compared to the ECMWF operational analysis data in the simulation period of 2-10 January 2012. Figure 3a shows the MBE (filled contours) and the RMSE (dotted contours) of U10 for the GWDSBL setup. Figure 3b show the difference of MBE (filled contours) and of RMSE (dotted contours) between the GWDSBL and LT (GWDSBL-LT). Figure 3b show the difference of MBE (filled contours) and of RMSE (dotted contours) between the GWDSBL and ST (GWDSBL-ST). For figures 3b and 3c positive values of the difference in RMSE shows improvement with the GWDSBL setup. For the difference of MBE negative values show improvement with the GWDSBL if the MBE of GWDSBL (3a) experiment is positive and deterioration if the MBE of GWDSBL is negative.

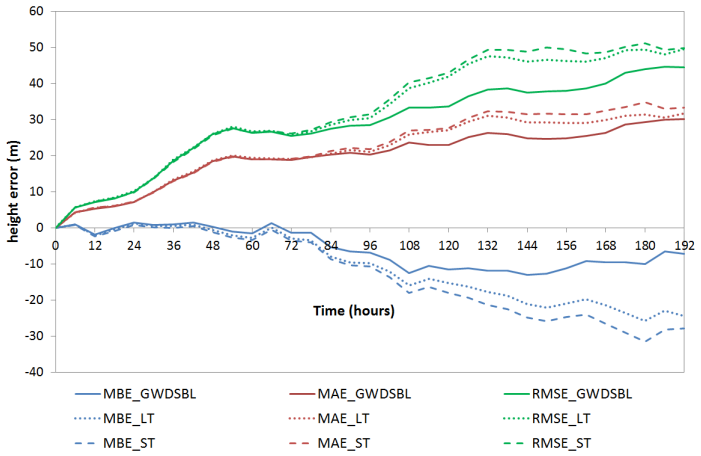


Figure 10. Temporal evolution of the mean statistical errors in the height of the 500 hPa field (averaged over all gridcells) for each of the model setups (GWDSBL, ST and LT).

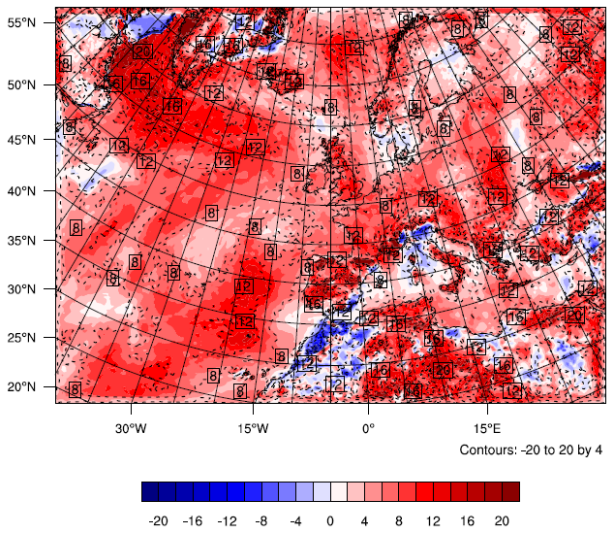
4.6. Relative humidity at the height of 2 meters

The last variable we investigate is the relative humidity at the height of 2 meters. An accurate representation of the RH_2m field can be crucial for forecasting phenomena like fog and stratocumulus clouds in the SBL. RH_2m is calculated for both the WRF model setup and the ECMWF operational analysis data.

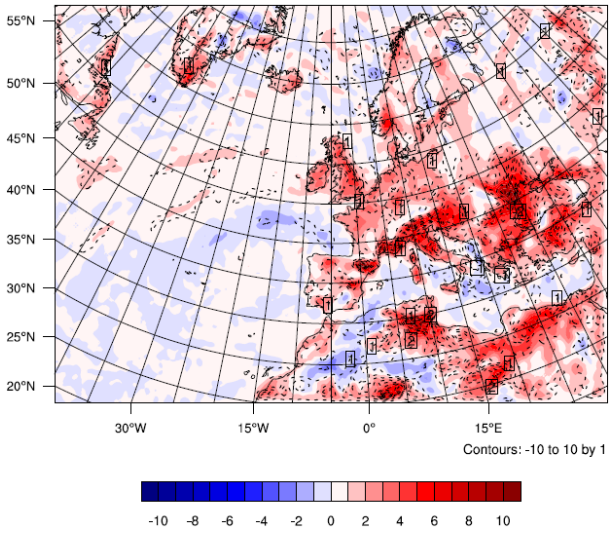
The GWDSBL has large positive bias in the calculation of the 2m relative humidity through the entire domain with only was positive bias regions (Figure 11a). The positive bias is ranging between 8% and 12% over Europe (8-9% over the whole domain, see Figure 12) and the Atlantic but there are also regions (eg. Sahara desert, west Greenland) with biases up to 20 %. For 2m temperature we found that GWDSBL run shows lower temperatures over the whole domain. Taking that into account we can hypothesis that the lower temperature of the whole domain can cause an increase in the relative humidity values by decreasing the saturation water vapour pressure, which results in lower saturation mixing ratio. Positive biases in surface pressure can have a similar effect resulting in further overestimation of RH_2m, while negative biases can result in underestimation. The combination of these to feedback mechanisms together with biases in the calculation of mixing ration at 2m are the main source of the relative humidity biases.

By comparing the errors of GWDSBL setup with the one from the LT setup we found that GWDSBL overestimates RH_2m by 1-3% over land (Figure 11b). On the contrary, above sea the LT and the GWDSBL shows only minor deviations. The extent of the differences can be seen also in the average over the whole domain statistical errors (Figure 12), where the MBE of GWDSBL is 1 % higher than that of LT. for MAE and RMSE different become even smaller (0.5%). Apart from the differences in 2m temperature (Figure 5b) and SLP (Figure 3b) that could have impacted RH_2m we also expect that there is an effect from 10m wind difference. The spatial patterns in Figure 11b shows remarkable similarity with the ones from Figure7b which leads us to conclude that near surface wind accuracy can have an effect in the accuracy of the mixing ratio scores and consequently the 2m relative humidity.

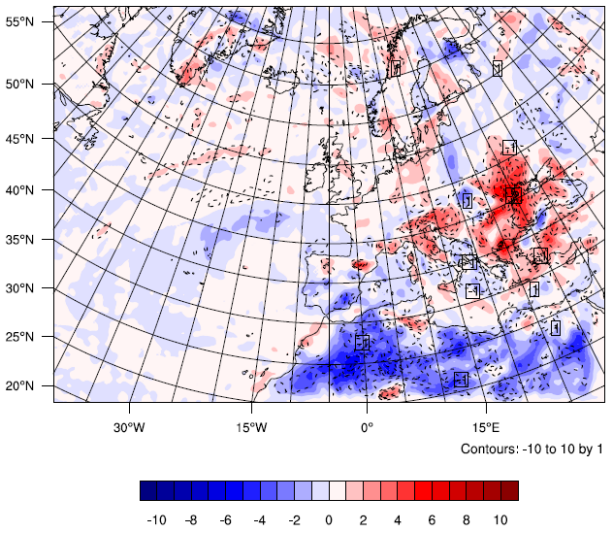
A similar comparison between GWDSBL and ST (Figures 11c and 12) reveals that the two setups have almost no difference when averaged over the whole domain but do show differences in their spatial patterns. GWDSBL show a smaller overestimation of RH_2m above north Africa compared to ST (1-2 %) and an higher one over Eastern Europe. This indicates the differences of RH_2m due to temperature differences between the setup can be offset by deviations in the SLP or 10m wind fields and vice versa resulting in near zero differences when averaged of the



(a)



(b)



(c)

Figure 11. Spatial distribution of mean statistical errors of RH at 2 meters compared to the ECMWF operational analysis data in the simulation period of 2-10 January 2012. Figure 3a shows the MBE (filled contours) and the RMSE (dotted contours) of U10 for the GWDSBL setup. Figure 3b shows the differences of MBE (filled contours) and of RMSE (dotted contours) between the GWDSBL and LT (GWDSBL-LT). Figure 3b shows the differences of MBE (filled contours) and of RMSE (dotted contours) between the GWDSBL and ST (GWDSBL-ST). For figures 3b and 3c positive values of the difference in RMSE shows improvement with the GWDSBL setup. For the difference of MBE negative values show improvement with the GWDSBL if the MBE of GWDSBL (3a) experiment is positive and deterioration if the MBE of GWDSBL is negative.

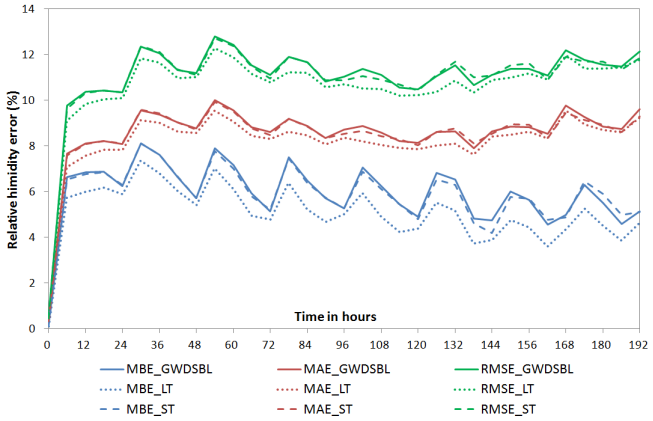


Figure 12. Temporal evolution of the mean statistical errors in 2m relative humidity (averaged over all gridcells) for each of the model setups (GWDSBL,ST and LT).

whole domain. Table 4 verifies our finding, regarding the small difference between GWDSBL and ST setup and difference up in the order of 1-2% between GWDSBL and LT setup, for the location for Cabauw tower.

Table 4. Statistical errors of 2m relative humidity (%) for all our WRF setups at the location of Cabauw tower. Comparison has been done against the observational data from Cabauw tower.

Experiment	MAE	MBE	RSME
GWDSBL	6.87	6.36	8.08
ST	4.95	3.66	6.13
LT	6.66	5.98	7.78

4.7. Taylor diagram

A summary of the model scores fo the run of 2-10 January including all 5 variables is presented in the Taylor diagram (Figure 13). In this diagram we compare the standard deviation and the spatial correlation of the 5 variables from the WRF setups with and the ECMWF operational analysis. The variables are bases on 2D fields averaged for the whole 8-day period. In this section we also included the table with the Taylor diagram skill score S (table 5) , which we calculate from equation 8.

The Taylor diagram follows well the qualitative results from the temporal and spatial analysis for all the variable in the case study of 2-10 of January 2012. For SLP the GWDSBL experiment shows the smallest difference from the ECMWF reference with a model skill score of 0.979 and is therefore in agreement with the finding in subsection 4.2. For the 2m temperature we found that GWDSBL has significant deviation from both LT and ST setups. This results supports our hypothesis that the deviations in 2m temperature between GWDSBL and the rest of the experimental setups in either due to the impacts of the dissipation of gravity wave in the SBL or due to the difference synoptic conditions. In section 4.3 we explain that the deterioration of T2m in the GWDSBL in an indicator of worse performance for this setups, but a results of increased 2m temperatures in LT and ST setups due to errors in near surface wind speeds and the synoptic conditions. For 10m wind speeds one can verify again the superiority of the new parametrization (skill score 0.882) compared to the LT and ST experiments (skill scores 0.841 and 0.849 respectively).

For relative humidity we find that the GWDSBL setup has the highest score (0.887), but will only small differences compared to LT and ST. The deterioration of the results compared to ST and LT setups could be attributed to certain differences between the setups. For the former, one should account the effects of the differences in 2m temperature, SLP and 10m wind speed (section

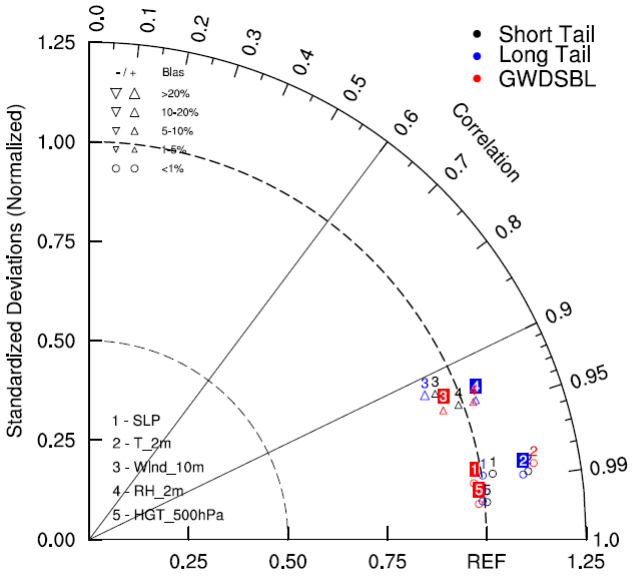


Figure 13. Taylor diagram for the model runs of 02/01 to 10/01 for all 3 model experiments, GWDSBL with red, LT with blue and ST with gray color. The diagram shows the score, compared to the reference ECMWF operational analysis data, for the 8 day averaged fields for sea level pressure (**1**), 2m temperature (**2**), 10m wind speed (**3**), relative humidity at 2m (**4**) and geopotential height at 500 hPa level (**5**). Axis x and y show the standardized deviations, which are normalized by the standard deviation of the ECMWF variable, and the hemispherical axis shows the correlation between the variable field for each experiment and the same field obtained as from the reference.

Table 5. Taylor diagram skill score S for all WRF setups and all variables in the model run 2-10 January. The skill score S represents the score of the model setup compared to the ECMWF operational analysis data, which are the treated as the point of reference. S score range from 1 to 0 (1 is the perfect score). Blue cell show lower skill score compared to GWDSBL and red cells higher.

Experiment	GWDSBL	LT	ST
SLP	0.979	0.974	0.973
T2m	0.956	0.969	0.964
U10m	0.882	0.841	0.849
RH2m	0.887	0.886	0.884
HGT_500 hPa	0.991	0.991	0.991

4.6). For the latter, one should also account the effects due to the difference in the enhanced mixing function. The height of the 500 hPa field is used as a more robust evaluation index for the score in the pressure fields and is closely related to the SLP scores (see section 4.5). The skill scores are very similar for GWDSBL, LT and ST setups, but we can still see a small improvement with the use of the GWDSBL scheme.

4.8. *Other case studies*

In this subsection we analyze all 16 of the of the model run with the use of the Taylor diagrams and the model skill score S (see eq 8). Our goal is to identify the number of the model runs in which the GWDSBL scheme outperforms LT and ST for all five variables (SLP, T2m, U10m, RH2m and HGT_500 hPa) that we use in our analysis. From our analysis of the Taylor diagram in section 4.7 we find that the skill score for each variable is in agreement with the corresponding spatial and temporal analysis of model accuracy. Therefore we have confidence that the skill score gives, foremost, a qualitative representation of the comparison between the three different model setups.

Regarding the 8-day averaged SLP field, we find a clear superiority of the GWDSBL over the LT and ST ones, with average S scores ranging between 0.91 and 0.995. GWDSBL has been able to outperform both LT and ST setups in 13 out of 16 model runs. In one model run ST outscores both GWDSBL and LT, while in the other 3 all three setups receive the same score. Over the first 8 model runs (run initialized between 15-29 of December) the difference between the skill scores (S) for SLP of the three setups are minimal (less than 0.01). The main reason for this similarity is that the synoptic conditions are dominated by high pressure systems, which all three model setups are able to represented accurately. On the other hands, model runs initialized between 31 December and 12 January have a strong activity of lower pressure systems over Europe. By averaging the skill score of SLP for all of the model runs, we found that the GWDSBL $\bar{S}$ score is 0.0042 higher than the LT one and 0.0073 higher than the ST setup (Figure 14). Under these conditions, GWDSBL is able to reproduce the SLP better than LT and ST due to a more accurate representation of the CCP and trajectory of individual systems.

The HGT_500 hPa Taylor diagram skill scores show similarities the score for SLP, but in this case S score range is smaller 0.980-0.999. We find that in 10 out of 16 model runs GWDSBL does better than LT and ST, 4 runs show no differences in S and in one model run ST scores higher. Moreover, the difference in the S scores have a similar pattern over the model runs with the ones from SLP. More specifically, all three setups have less than 0.001 difference in the first 7 model runs and higher difference for the rest (with the exception of model run 10, 2-10 January). The averaged skill score $\bar{S}$ (see Figure 14) for the GWDSBL runs is 0.0008 and 0.0012 higher than that of the LT and ST setup respectively. The almost identical behaviour of the SLP and HGT_500 hPa scores verifies that these two variables are closely connected, but the latter is much less sensitive to deviations from the reference synoptic conditions compared to the former one.

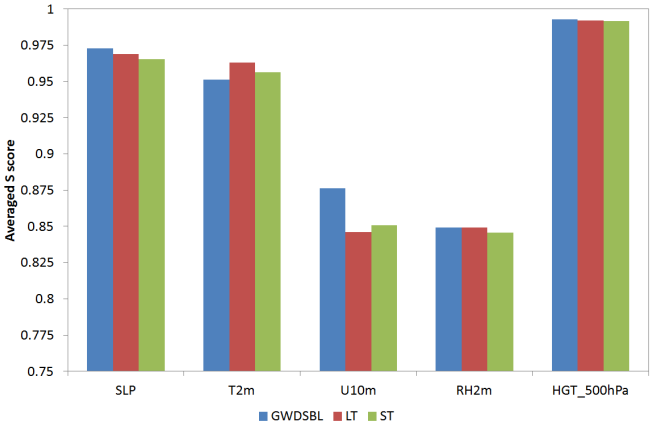


Figure 14. Taylor diagram skills scores S averaged over the 16 model runs for all variables. The averaged skill score $\bar{S}$ represents the score of the model setup compared to the ECMWF operational analysis data, which are the treated as reference. S score range from 1 to 0 (1 is the perfect score).

Skill scores for 10m wind show a remarkable superiority of the new parameterization. The model skill scores for 10m wind range between 0.81 and 0.915, with the GWDSBL scoring on average 0.03 to 0.025 higher than LT and ST respectively (Figure 14). The improvement originates from the reduction of near surface wind in the GWDSBL run due to enhanced momentum drag from the dissipation of gravity waves. We can verify this, because the difference are visible even in model runs with no difference in the synoptic conditions. Furthermore, the small improvement (less than 0.01 of the S score) with the use ST scheme over the LT indicates that the improvement in near surface wind due to the difference in the mixing function is not as important as the improvement introduced with the new parameterization.

For the 2m temperature we can observe the exact opposite behaviour regarding the performance of the GWDSBL scheme. Skill scores range between 0.941 and 0.972. The GWDSBL setup

obtains the lowest skill scores for all of the 16 model runs, with LT obtaining the highest ones and ST lying in between. The difference in the averaged skill score (see figure 14) between GWDSBL and LT is approximately 0.015 and it is reduced to 0.005 between GWDSBL and ST. The differences in the skill scores are relatively stable, even in model runs with small differences in the synoptic weather. Therefore, we can argue that the differences in 2m temperatures skill score are either due to difference in mixing function (between LT and ST, GWDSBL) or due to weaker mechanical mixing as a result of smaller 10m wind speed in the new parametrization (between GWDSBL and ST).

Finally we analyses the skill score for the 2m relative humidity. The three model setups receive skills scores between 0.805 and 0.888. The LT setup outscores GWDSBL and ST one in 13 out of 16 model runs, while GWDSBL outscores LT in the rest. GWDSBL has also higher skill score in 13 model run compared to the ST setup. The difference between the model setups are highly variable and dependent on the synoptic conditions. In model runs dominated by high pressure systems, the LT setups show larger difference compared to the rest of the model. On the other hand, in model run with larger amount of strong low pressure systems all three model setups have generally high and indifferent skill scores. On average LT setups scores 0.004 higher than GWDSBL, the latter one has a 0.035 increase in $\bar{S}$ compared to the ST setup (see Figure 14).

5. Discussion

For this study we use the parametrization for small-scale orographic drag introduced by the study of Steeneveld *et al.* (2008) and adapted for the WRF-ARW 3D model. This parametrization is based on simplified physics and it does not resolve the gravity waves fields directly, but it only represents the impact of gravity wave dissipation and saturation in the momentum drag of the SBL. Steeneveld *et al.* (2008) explained that this approach of quantifying the gravity wave drag is based on the assumption that the wind speed and stratification with height is sufficiently small (i.e Wentzel–Kramers–Brillouin assumption) and the wind profile has a hyperbolic tangent shape. These two assumptions can be unrealistic in parts of the SBL, with strong wind shear and large temperature gradients (Balsley *et al.* 2003). For the calculation of the subgrid-scale orographic amplitude and the wavenumber we use the theory introduced by Kim and Doyle (2005), which might be a simplification of the subgrid-scale orography.

The way we calculate the gravity wave drag in the GWDSBL parametrization is based on equation 4, which is an approximation of equation 3 for weak background winds Nappo (2002). In his study Steeneveld *et al.* (2008) shows that using the linear an approximation of the wave drag instead of the non-linear one provides similar results. In one of our sensitivity tests we parametrize the gravity wave drag using equation 3 and compare it with the linear an approximation. We found that there is no significant difference between the two.

Gravity wave dissipation does not have only effects on the momentum drag. Nappo (2002) showed that gravity waves can also impact pressure, vertical wind and temperature. Indeed the effects of gravity wave dissipation in the vertical mixing of heat has been verified by the study of Fritts and Dunkerton (1985), who show that gravity waves breaking can result in counter gradient of potential temperature and therefore a source of convective instability. Implementation of this mechanism in the SBL can result in an increase of vertical mixing of heat, which can results in a reduction of the negative bias of 2m temperature that we find in the GWDSBL experiment.

In section 3.2 we explain the procedure of removing the systematic errors, which are visible from time step zero for the SLP and HGT_500 hPa variables. While we are confident that the initial errors are the same for all model setups, we are not entirely certain that the extent of the systematic errors is exactly the same at all time steps. Therefore we calculated the MBE and RMSE difference maps (eg. Figures 3b and 3c) in both cases, one with the systematic error removed from each time step and one in which we include it. We have found no difference in the figures produced with the two different methods. Only the initial figures of MBE and RMSE for GWDSBL are altered (eg. Figure 3a) due to the exclusion of the initial error.

State of the art NWP models have a much higher vertical resolution compared to the 34 levels that we are using. To investigate the impacts of a higher vertical resolution in the model performance we did a sensitivity test including 61 vertical level in the model setup for the case of 31 December 2011 to 8 January 2012. We found that the increase in vertical resolution provides a higher accuracy in the cyclonic life cycle and a reduction in the overestimation of high pressure systems intensity. Unfortunately the increases in computational time and data output lead us to use only 34 vertical levels.

For the calculation of the τ_{wave} in equation 4 U is calculated as wind speed at the top of the boundary layer. This wind speed is also used as a criterion for identifying if the wave reach the critical point of wave breaking ($N/U \leq k_s$). Knowing that the SBL is usually characterized by significant gradient of wind with height (Balsley *et al.* 2003), can result in an invalidity of our hypothesis for cases where the subgrid-scale orography is very small compared to the boundary-layer height or in cases with strong stratification of the wind in the SBL. In order to test the validity of using hypothesis is valid we have conducted a sensitivity test using the mean wind speed in the SBL instead of the one on the top U_{hpbl}. We found that while there is a small deviation from the default case, due to small values of the gravity wave drag as a result of weaker background wind, there is no significant different in the overall performance between the two experiments.

In this study we use a K-profile function (Holtslag *et al.* 1990; Holtslag and Boville 1993). Another possibility is to use a expression for K_m based on the local stability parameters, a method that was used in the study of Steeneveld *et al.* (2008) and its also used by ECMWF (Sandu *et al.* 2013). In one of our sensitivity test we compared the two different methods and we found that the first one gives slight better results, especially under weakly stable conditions. The second approach takes into account the effect of wind shear on the diffusivity. As a result under very stable conditions, with strong stratification of the wind, there will be more mechanical mixing in the SBL. This mechanism can possibly reduce the biases in 2m temperature in regions with heavy stratification (see 4.3). A interesting topic for research will be to study if this mechanism can indeed improve 2m temperature score without deteriorating the performance for the rest of the variables in the majority of the model runs.

The LT experiment in our study is based on the mixing function in equation 2. This "long-tail" mixing function is a weaker compared to the one function of Louis (1979) and the revised one by Viterbo *et al.* (1999) that have been used by ECMWF. The difference is that the LT function results in higher values for the momentum diffusion (F_m) for values of $Ri < 0.2$ and smaller ones above that threshold. These difference inhibit us from making concrete comparison with the ECMWF forecast model. This obstacle could be avoided if one uses the local mixing approach for K_m. Connecting this with topic the previous paragraph we advice the repetitions of this study with the use of the local method for calculating the diffusion coefficient under stable conditions.

6. Conclusions

In this paper we investigate the role of small-scale orographic gravity wave drag as a possible mechanism to increase momentum drag within the SBL. We also analyze the impact of this mechanism in the life cycle of synoptic systems and the near surface meteorology. In our hypothesis we have assumed that the representation of the momentum drag due to dissipation of gravity wave is a possible mechanism to bridge the cap in the momentum drag between the a "short-tail" and a "long-tail" mixing functions and it can improve the scores for synoptic system and SBL variables. To realize this hypothesis we have parametrize the enhanced momentum drag induced by gravity wave of small subgrid-scale orography and we have add this to the turbulent drag calculated by a "short-tail" mixing function. Our goal is to identify whether this new parametrization (GWDSBL) can function similarly to a "long-tail" regarding the score for the representation of synoptic systems and 2m temperature, while maintaining the accuracy of a "short-tail" mixing function in the representation of near surface wind and SBL height.

The updated scheme (GWDSBL) has been tested against a "short-tail" mixing function scheme (ST) and a "long-tail" mixing function scheme (LT) for sixteen 8-day forecasts, with re-initialization of the forecast every 48 hours. This approach allows us to test the three model setups for more than a whole month (15 December 2011 to 22 January 2012) during winter. In his paper we emphasize in the result of a case study initialized the 2nd of January, but we also refer to statistics and findings from the whole study period and the sensitivity analysis on the GWDSBL setup. For verification of the three model setups we use the ECMWF operational analysis and observations from the Cabauw weather station.

For the case study of 2-10 January 2012 the new parametrization (GWDSBL) has been able to reproduce the cyclonic core pressure of two low pressure systems with higher accuracy compared to LT (1% reduction in RMSE for cyclone 1 and 50% for cyclone 2) and ST (16% reduction in RMSE for cyclone 1 and 50% for cyclone 2) setup. Regrading the life cycle of cyclone 2, GWDSBL even manages to outscore ECMWF forecast for the CCP, by 2.07 hPa in RMSE. The improvement in the representation of synoptic conditions has resulted in better score for SLP in the GWDSBL setup. RMSE over the whole domain is reduced up to 1 hPa near the end of the simulation, while MBE is improved even more (1.5 hPa). Improvements also occurs in the height of the 500 hPa fields compared to ST and LT setups.

Regarding near surface wind we find a reduction in the negative wind bias by up to 0.4 ms^{-1} (50 %) and also for the other statistical scores when GWDSBL is used. Despite these improvements we found a small deterioration of 2m temperature scores (up to 0.3 K) compared to the ST and LT runs, which could be a result of different synoptic conditions between GWDSBL or stronger mixing due to higher wind speeds in the other two model setups.

Finally, we have analyzed all 16 model run, which cover the period of 15 December 2011 to 22 January 2012. Our findings verify that even in the cases when GWDSBL performs poorly, SLP and HGT_500 hPa scores are almost always higher than the ST and LT setups. Moreover, near surface wind speeds are always better represented when we include the new parametrization for gravity wave drag. Summarizing all our findings we are confident in stating that the updated parametrization is able to function effectively as a "long-tail" under the right synoptic conditions.

7. Acknowledgements

This study has been supported by Wageningen University and the Meteorology and Air Quality research group of Wageningen University, who provided the facilities necessary for the realization of the research. The authors also want to thank ECMWF for the initial and boundary conditions for the WRF-3D model runs. The first author want to thank the two co-authors and supervisors, G.J. Steeneveld and A.A.M Holtslag, for their support and guidance throughout the whole part of the research process and also thank Kees van den Dries for his support with setting up the WRF-3D model runs.

References

Anthes RA, Keyser D. 1979. Tests of a fine-mesh model over europe and the united states. *Monthly Weather Review* **107**(8): 963–984.

Balsley BB, Frehlich RG, Jensen ML, Meillier Y, Muschinski A. 2003. Extreme gradients in the nocturnal boundary layer: Structure, evolution, and potential causes. *Journal of the Atmospheric Sciences* **60**(20): 2496–2508.

Beare RJ. 2007. Boundary layer mechanisms in extratropical cyclones. *Quarterly Journal of the Royal Meteorological Society* **133**(623): 503–515.

Beljaars A, Viterbo P. 1998. *Role of the boundary layer in a numerical weather prediction model. clear and cloudy boundary layers*. Royal Netherlands Academy of Arts and Sciences: Amsterdam, pp. 287–304.

Beljaars ACM, Holtslag AAM. 1991. Flux parameterization over land surfaces for atmospheric models. *Boundary-Layer Meteorology* **118**(3): 327–341.

Boutle IA, Beare RJ, Belcher SE, Brown AR, Plant RS. 2009. The moist boundary layer under a mid-latitude weather system. *Boundary-Layer Meteorology* **134**(3): 367–386.

Boutle IA, Belcher SE, Plant RS. 2015. Friction in mid-latitude cyclones: an ekman-pv mechanism. *Atmospheric Science Letters* **16**(2): 103–109.

Cassano JJ, Higgins ME, Seefeldt MW. 2011. Performance of the weather research and forecasting model for month-long pan-arctic simulations. *Monthly Weather Review* **139**: 3469–3488.

Chen F, Dudhia J. 2001. Coupling an advanced land surfacehydrology model with the penn statencar mm5 modeling system. part i: Model implementation and sensitivity. *Monthly Weather Review* **129**: 569–585.

Chimonas G, Nappo CJ. 1989. Wave drag in the planetary boundary layer over complex terrain. *Boundary-Layer Meteorology* **47**(1).

Choi HJ, Hong SY. 2015. An updated subgrid orographic parameterization for global atmospheric forecast models. *Journal of Geophysical Research: Atmospheres* **120**(24): 12 445–12 457.

Cuxart J, Holtslag AAM, Beare RJ, Bazile E, Beljaars A, Cheng A, Conangla L, Ek M, Freedman F, Hamdi R, Kerstein A, Kitagawa H, Lenderink G, Lewellen D, Mailhot J, Mauritsen T, Perov V, Schayes G, Steeneveld GJ, Svensson G, Taylor P, Weng W, Wunsch S, Xu KM. 2006. Single-column model intercomparison for a stably stratified atmospheric boundary layer. *Boundary-Layer Meteorology* **118**(2): 273–303.

Dudhia J. 1989. Numerical study of convection observed during the winter monsoon experiment using a mesoscale two-dimensional model. *Journal of the Atmospheric Sciences* **46**: 3077–3107.

Dyer AJ. 1974. A review of flux-profile relationships. *Boundary-Layer Meteorology* **7**(3): 363–372.

Fritts DC, Dunkerton TJ. 1985. Fluxes of heat and constituents due to convectively unstable gravity waves. *Journal of the Atmospheric Sciences* **42**(6): 549–556.

Gilliam RC, Pleim JE. 2006. Performance assessment of new land surface and planetary boundary layer physics in the wrf-arw. *Journal of Applied Meteorology and Climatology* **49**(4): 760–774.

Holton J. 1992. *An introduction to dynamical meteorology*. Academic Press: New York, USA.

Holtslag AAM, Boville BA. 1993. Local versus nonlocal boundary-layer diffusion in a global climate model. *Journal of Climate* **6**(10): 1825–1842.

Holtslag AAM, Bruijn EIFD, Pan HL. 1990. A high resolution air mass transformation model for short-range weather forecasting. *Monthly Weather Review* **118**(8): 1561–1575.

Holtslag AAM, Svensson G, Baas P, Basu S, Beare B, Beljaars ACM, Bosveld FC, Cuxart J, Lindvall J, Steeneveld GJ, Tjernstrom M, Wiel BJHVD. 2013. Stable atmospheric boundary layers and diurnal cycles: Challenges for weather and climate models. *Bulletin of the American Meteorological Society* **94**(11): 1691–1706.

Holtslag B. 2006. Preface: Gewex atmospheric boundary-layer study (gabls) on stable boundary layers. *Boundary-Layer Meteorology* **118**(2): 243–246.

Hong SY, Choi J, Chang EC, Park H, Kim YJ. 2008. Lower-tropospheric enhancement of gravity wave drag in a global spectral atmospheric forecast model. *Weather and Forecasting* **23**(3): 523–531.

Hong SY, Dudhia J, Chen SH. 2004. A revised approach to ice microphysical processes for the bulk parameterization of clouds and precipitation. *Monthly Weather Review* **132**: 103–120.

Hong SY, Noh Y, Dudhia J. 2006. A new vertical diffusion package with an explicit treatment of entrainment processes. *Monthly Weather Review* **134**(9): 2318–2341.

Jimenez PA, Dudhia J, Gonzlez-Rouco JF, Navarro J, Montvez JP, Garca-Bustamante E. 2012. A revised scheme for the wrf surface layer formulation. *Monthly Weather Review* **140**(3): 898–918.

Kim YJ, Arakawa A. 1995. Improvement of orographic gravity wave parameterization using a mesoscale gravity wave model. *Journal of the Atmospheric Sciences* **52**(11): 1875–1902.

Kim Yj, Doyle JD. 2005. Extension of an orographic-drag parametrization scheme to incorporate orographic anisotropy and flow blocking. *Quarterly Journal of the Royal Meteorological Society* **131**(609): 1893–1921.

Kleczek MA, Steeneveld GJ, Holtslag AAM. 2014. Evaluation of the weather research and forecasting mesoscale model for gabls3: Impact of boundary-layer schemes, boundary conditions and spin-up. *Boundary-Layer Meteorology* **152**(2): 213–243.

Lapworth A. 2015. Observations of the site dependency of the morning wind and the role of gravity waves in the transitions. *Quarterly Journal of the Royal Meteorological Society* **141**(686): 27–36.

Lapworth A, Claxton BM, McGregor JR. 2015. The effect of gravity wave drag on near-surface winds and wind profiles in the nocturnal boundary layer over land. *Boundary-Layer Meteorology* **156**(2): 325–335.

Lilly DK. 1972. Wave momentum flux-a garp problem. *Bulletin of the American Meteorological Society* (53): 17–23.

Louis JF. 1979. A parametric model of vertical eddy fluxes in the atmosphere. *Boundary-Layer Meteorology* **17**(2): 187–202.

Mauritsen T, Svensson G. 2007. Observations of stably stratified shear-driven atmospheric turbulence at low and high richardson numbers. *Journal of the Atmospheric Sciences* **64**(2): 645–655.

McCabe A, Brown AR. 2006. The role of surface heterogeneity in modelling the stable boundary layer. *Boundary-Layer Meteorology* **122**(3): 517–534.

Mlawer EJ, Taubman SJ, Brown PD, Iacono MJ, Clough SA. 1997. Radiative transfer for inhomogeneous atmospheres: Rrtm, a validated correlated-k model for the longwave. *Journal of Geophysical Research: Atmospheres* **102**: 16 663–16 682.

Nappo CJ. 2002. *An introduction to atmospheric gravity waves*. Academic Press:London, pp. 47–84.

Palmer TN, Shutts GJ, Swinbank R. 1986. Alleviation of a systematic westerly bias in general circulation and numerical weather prediction models through an orographic gravity wave drag parametrization. *Quarterly Journal of the Royal Meteorological Society* **112**(474): 1001–1039.

Plant RS, Belcher SE. 2007. Numerical simulation of baroclinic waves with a parameterized boundary layer. *Journal of the Atmospheric Sciences* **64**: 4383–4399.

Sandu I, Beljaars A, Bechtold P, Mauritsen T, Balsamo G. 2013. Why is it so difficult to represent stably stratified conditions in numerical weather prediction (nwp) models? *Journal of Advances in Modeling Earth Systems* **5**(2): 117–133.

Shin HH, Hong SY, Dudhia J, Kim YJ. 2010. Orography-induced gravity wave drag parameterization in the global wrf: Implementation and sensitivity to shortwave radiation schemes. *Journal of Applied Meteorology and Climatology* **2010**: 1–8.

Skamarock W, Klemp J, Dudhia J, Gill D, Barker D, Duda M, Huang XY, Wang W. 2008. A description of the advanced research wrf version 3. Technical report, NCAR Technical Note NCAR/TN-475+STR.

Steeneveld GJ. 2014. Current challenges in understanding and forecasting stable boundary layers over land and ice. *Frontiers in Environmental Science* **2**(41).

Steeneveld GJ, Holtslag AAM, Nappo C J van de Wiel BJH, Mahrt L. 2008. Exploring the possible role of small-scale terrain drag on stable boundary layers over land. *Journal of Applied Meteorology and Climatology* **47**(10): 2518–2530.

Stensrud D. 2007. *Parameterization schemes:keys to understanding numerical weather prediction models*. Cambridge University Press: Cambridge, pp. 138–184.

Svensson G, Holtslag AAM. 2009. Analysis of model results for the turning of the wind and related momentum fluxes in the stable boundary layer. *Boundary-Layer Meteorology* **132**(2): 261–277.

Taylor KE. 2001. Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres* **106**(D7): 7183–7192.

Troen IB, Mahrt L. 1986. A simple model of the atmospheric boundary layer; sensitivity to surface evaporation. *Boundary-Layer Meteorology* **37**(1): 129–148.

Viterbo P, Beljaars A, Mahfouf J, Texeira J. 1999. The representation of soil moisture freezing and its impacts on the stable boundary layer. *Quarterly Journal of the Royal Meteorological Society* **155**(559): 2401–2426.

Walsh JE, Chapman WL, Romanovsky V, Christensen JH, Stendel M. 2008. Global climate model performance over alaska and greenland. *Journal of Climate* **21**(23): 6156–6174.

Willmott CJ. 1982. Some comments on the evaluation of model performance. *Bulletin of the American Meteorological Society* **63**(11): 1309–1313.

