Response letter

Disentangling the dynamic interplay between affect fluctuations and depressive complaints using network analysis.

Comments editor:

1. I struggled to understand the justification for the study, and would suggest that the introduction be greatly expanded. The hypotheses are relatively general ("how does affect relate to the evolution of depressive symptoms over several months"), and the specific sub-aims (COVID links, individual differences in trajectories, and mechanisms) aren't justified at all. Even if these are exploratory, it would help readers to understand why these specific questions were the focus of the study. I also found it distracting that some additional aims(such as "Do affect trajectories differ across "types" of depression trajectories?") weren't raised in the introduction. Although one might argue that they are a specific instantiation of the briefly introduced hypotheses in the introduction, there's nothing in the introduction that justifies that specific hypothesis (as R2 notes: why group individuals rather than treat depression as a continuous moderator/predictor?).

We would like to thank the editor for pointing out that the introduction needed a large expansion on the justification and explanation of our research topic. We completely agree with this comment and we have therefore greatly expanded the introduction. We paid much attention to the justification of our study, namely, to investigate whether we can better understand the heterogeneity of depression complaints by studying the relationship between affect regulation and the evolution of depressive complaints within a context of prolonged stress (see page 3-6). We have specifically focussed on explaining the setting of this study during COVID-19 (see page 5-6), why we would expect individual differences in affect regulation to play a role (see page 3-4), and the mechanisms that we aim to investigate (see page 6-7) – in line with your valuable suggestions. We also pay more attention to outlining our research aim and sub-aims along the different methods we will use in the introduction. The hypotheses are now more specific, following the aims and expectations from our pre-registration (see page 6-7). As such, the introduction is now streamlined with the 'statistical analyses' paragraph in the Methods section (see page 9-13).

2. I would like to see psychometric information provided for the measures.

Thank you for pointing this out, we agree that this information should be provided and have now added the psychometric information about the short version of the PANAS and the PHQ-9 in the manuscript in the following manner (page 8):

"The reliability of both PA and NE short-form scales is high, with a Cronbach alpha of .78 for the PA scale and .87 for the NE scale (Mackinnon et al., 1999)."

"The internal reliability of the PHQ-9 is high, with a Cronbach's alpha ranging between .86-.89 (Kroenke et al., 2001)."

3. I was confused by citing Jordan as a justification for minimum within person observations for MLMs; that paper is about networks and reports a minimum of 20 for mlVARs. Was this perhaps a typo?

We thank the editor for noting this typo. We were indeed referring to the multilevel VAR model, not to MLM. We have changed the typo accordingly (see page 9):

"Following the recommendations for estimating a multilevel VAR model (Jordan, Winer & Salem, 2020), we selected participants who completed surveys for at least 20 measurement occasions, resulting in a final sample size of N=228 participants (16.8%)."

4. I don't necessarily object to using mlVAR for estimating the interrelations among affect and depressive complaints, but I don't think the analyses are clearly or specifically justified. Why not use, for example, a multilevel model? The specific analysis should be more clearly justified in terms of the motivating question at hand.

We apologise for not making sufficiently clear why we have used a multilevel VAR network model and now explicitly explain our reasons for using a multilevel VAR network model instead of a multilevel model, namely, that with a network (multilevel VAR) model we can deconstruct the concepts of positive affect, negative affect, and depressive complaints into smaller elements that are simultaneously independent and dependent variables. In a longitudinal context, this gives us the opportunity to look at how fluctuations in specific elements of negative and positive affect are related to evolutions of certain mental health complaints. (See our revised introduction; page 4):

"The possible dynamic mechanisms underlying the relation between affect and the development of depressive complaints can be investigated using network analysis (Hoorelbeke, van den Bergh, Wichers & Koster, 2019). Network models estimate the interrelations between variables on a detailed level, in which concepts such as PA and NA can be deconstructed into smaller elements (e.g., feeling inspired or excited for PA; feeling afraid or upset for NA). This means that all smaller components of PA, NA and distinct depressive complaints are both dependent and independent variables simultaneously (Ryan, Bringmann & Schuurman, 2022). With longitudinal networks that are estimated from data

including assessments of PA, NA, and depressive complaints over a longer stressful period from multiple participants, we can look into various directed pathways of how fluctuations in negative and positive affect are related to evolutions of specific mental health complaints. In this way, we aim to deconstruct the causal relationship between affect and depressive complaints and take a step further in understanding the heterogeneity of depression."

5. I also find both strengths and weaknesses in the temporal scaling of this study. On the one hand, I think it's a strength to study these associations over a roughly 3 month period. On the other, having measurements at 3 days apart makes some assumption that affect on a Tuesday will impact depressive symptoms on a Friday. I don't think this is completely ridiculous, but I think this temporality has to be specifically acknowledged in the manuscript. For example, this might explain differences in the temporal results vs. the within/between results.

We thank the editor for raising this important point and apologise for not addressing this in the previous version of our manuscript. We now acknowledge the temporal scaling of our study and discuss its implications on how to interpret the network structure in the result section (see page 17-18):

"It is important to note that questions regarding affect and depressive complaints were prompted differently. Affect was prompted: "Indicate which description best describes how you currently feel, right now in the moment" while questions regarding depressive complaints were prompted: "In the last several days, how often have you been bothered by any of the following problems". This difference in phrasing, as well as the temporal scaling of our study (roughly 3 days between assessments) influences the interpretation of the estimated networks. The relations in the average contemporaneous network are a reflection of how depressive complaints over the days prior to the assessment and affect at the moment of assessment co-occur (e.g., how depressive complaints over a period of 3 days, let's say monday through wednesday, co-occurs with affect on wednesday). Relations in the average temporal network are a reflection of how depressive complaints over the days prior to the assessment predicts affect at the next assessment occasion (e.g., how depressive complaints over a period of 3 days, for example, monday through wednesday, predicts affect on friday) and vice versa. The relations in the between-person network are a reflection of how, on average, depressive complaints relate to affect."

In addition, we refer back to the temporal scaling of the study in the discussion, see page 24:

"*Third, affect and depressive complaints may operate on different timescales. To the best of our knowledge, there are no current techniques available that can account for these possible time-scale differences in processes in the field of network analysis (Bringmann et al., 2021). However, the time-scale differences in affect and depressive complaints where taking into account in the questionnaire.*

Questions regarding affect were prompted: "Indicate which description best describes how you currently feel, right now in the moment" while questions regarding depressive complaints were prompted: "In the last several days, how often have you been bothered by any of the following problems". This difference in phrasing captures the idea that affect operates on a faster time-scale (e.g., fluctuates within a day), while depression operates on a slower time-scale (e.g., fluctuates from day-to-day)."

6. Given the original sample size, I think it's important to report whether or how included participants differed from those excluded

We agree with the editor that it is important to report whether and how the included participants differed from those excluded. We have now included sample characteristics (see Table 1; page 14-15, and Table B1; page 39-40 in the manuscript on page), and means and standard deviations of the variables of interest (see Table 2; page 15, and Table B2; page 40 in the manuscript page) for both the original sample and the selected sample. For the original full sample these tables have been added to a supplement, for the selected sample the tables have been added to the manuscript itself. In addition, we acknowledge the differences between the original sample and the selected sample when discussing our results (see page 14):

"Our sample (n=228) differed from the original full sample (n=1355) in their mean PHQ-9 score at baseline; the original full sample had a higher PHQ-9 mean at baseline. Our sample did not differ in their PANAS scores at baseline from the original full sample (n=1355), see Supplement B, Table B1 for more details on sample characterization, means standard deviations, and range for the PHQ-9 and PANAS for the original full sample."

and in the discussion (see page 25-26):

"Second, due to the high amount of missing data in the original full sample, we selected participants that answered enough measurement occasions. The selected sample of participants who completed at least 20 measurement occasions had a lower PHQ-9 mean at baseline than the original full sample. Therefore, it can be that we may have missed more severe depression cases. At the same time, our selected sample still included 129 (56.6%) participants who scored above the cut-off of severe depressive complaints at baseline and 41 participants (18.0%) who consistently scored above the cut-off for severe depressive complaints."

7. I was also confused by the lack of reporting on other missing data. I presume that there was missing data for included participants; what was it related to? What was the extent of it, and how was it handled in analysis?

We apologise for any confusion regarding our reporting of missing data. It is important to note that for all participants included in the current analyses, they completed at least 20 measurements (as explained in the method and result section, see page 9 and page 13). Within each of these measurements, there is no missing data (i.e., there is no within wave missingness). That is, for each registered response we always have complete data on both the PHQ-9 and the PANAS as these were implemented in the same questionnaire with forced responses on all questions. The average of 28% missing data for each included participant thus relates to complete missed assessments between their first completed questionnaire and their last completed questionnaire. When omitting these assessments, we retained at least 20 assessments per participant following our inclusion criteria. We explain this in the result section:

"Over the course of four months, on average, participants completed 23±3 (mean±SD) assessments. On average, 9±3(mean±SD) assessments were missing, thus, on average, 28% waves of data are missing per individual. As responses for all variables of interest where forced, there is no within wave missingness. For missing data patterns per participant, see Figure A3 in Supplement A."

As well as in Supplement A page 35-37:

"Each participant has completed data for a minimum of 20 and a maximum of 33 measurement occasions. On average, participants had 9.4 (SD = 2.7) missed measurement occasions. Thus, on average, 28% of a participant's data is missing. It is important to note that within each of the measurement occasions, there is no missing data. That is, for each registered response, data on both the PHQ-9 and the PANAS are completed as these were implemented in the same questionnaire with forced responses on all questions. Thus, the average of 28% missing data, relates to complete missed assessments between participants' first questionnaire and their last completed questionnaire. See Figure A3 for missing data patterns per participant."

8. I think it would also help to better justify the use of loess to estimate trajectories. The downside is that this was done at essentially the individual level, and then aggregated by a-priori groupings of depressive symptoms. This seems problematic to me because there's no acknowledgement of error, either in the grouping of symptom patterns, or of the relations between symptoms and affect trajectories. It strikes me that other methods, like latent growth curve, mixture/class, or multilevel models, might be able to provide much of the sample information while also providing point estimates and estimates of uncertainty. This ad-hoc

method as provided in the present study feels descriptive at best, and although that may be reasonable, it should be justified.

We thank the editor for raising this important point. Upon rereading we realise that the way in which we presented these analyses was not properly justified. We agree that the way in which our rationale was presented now, another method such as those presented by the editor would have been more appropriate. However, our objective indeed was to just describe different trajectories in pre-defined groups of depression evolution. Descriptively, our objective was to explore whether the affect trajectories would differ in different pre-defined groups of depression evolution (i.e., aggravation or alleviation, and consistently no, mild, or severe complaints). We are now more explicit about this descriptive objective and have adjusted our justification for using loess accordingly (page 6):

"To elucidate the working mechanisms behind the interplay between affect and depressive complaints we first aim to explore whether the evolution of depressive complaints in the beginning of the COVID-19 pandemic (i.e., worsening or alleviation of complaints) is accompanied by general trends in affect fluctuations over that same period. Since we have scores on depressive complaints throughout the studied period, we divide the sample into sub-groups of people who experienced a meaningful change during the studied period, either an aggravation or alleviation of complaints, and sub-groups whose scores can consistently be classified as no complaints, mild complaints, or severe complaints. We then visually inspect the trajectories of positive and negative affect during that same period. We expect that people who experienced an increase in depressive complaints would show high levels of NA, whereas people who reported a decline in complaints would show high levels of PA. Following our hypotheses that the evolution of depressive complaints may be linked to differences in affect regulation, we then explore the affect trajectories for people who experience a meaningful change in their depressive complaints over the studied period."

Review 1:

1. The authors cite Molenaar (2004) but make no efforts to test whether and to what extent their time series data violate any of the ergodic assumptions (stationarity, homogeneity, memorylessness). If these assumptions are violated a difference between individual and group trajectories is expected, if they are not violated, they should be approximately the same, that is what the ergodic condition implies. The authors seem to assume ergodicity was indeed violated, because in the discussion they note time varying VAR models would have been preferred, but could not be used with their data… I believe it is appropriate to ask the authors to add some comments on the properties of their time series data that inform the reader about nonstationarity, heterogeneity and the presence of long range temporal correlations beyond lag-1. Also, what are the consequences of violating these assumptions for their conclusions? For an example of testing assumptions see Olthoff et al., Complexity in psychological self-ratings:Implications for research and practice BMC Medicine, 2020; for an example of consequences of violating the assumptions, see Fisher et al.Lack of group-to-individual generalizability is a threat to human subjects research PNAS 2018.

We would like to thank the reviewer for this important point, and have greatly rewritten the introduction. We agree with the reviewer that the former version put quite some emphasis on the affect state vs. trait discussion, and in that context, the ongoing debate in psychology on inferring things at the individual level from population data while psychological processes are rarely ergodic. We removed this section and the citations related to those debates (e.g., the paper by Molenaar 2004), as our real goal was more conceptual, namely, to investigate whether we can better understand the heterogeneity of depression by studying the relation between the evolution of depressive complaints and affect regulation in detail (see page 4-7). Our current introduction setting is almost completely rewritten and expanded, to better explain our research aim. We apologise for the confusion by putting too much focus on the ergodicity discussion in our former version. Currently, we believe the suggested checks are no longer in line with the new research aims and sub-aims as written in the novel introduction (see page 6-7). However, if the reviewer thinks the suggested checks are still relevant after reading our new manuscript, we would be glad to do these analyses.

2. The authors construct the mlVar networks and subsequently find that the 3 expected clusters are absent from the temporal mlVar network. They explain: “This suggests that the relations among affect and depressive complaints within persons over time may be substantially different from their relations within one timepoint or between-persons. ”This is re-stating the result, but what does it actually mean for the conclusions about the relationship between the 3 assumed clusters? What does it mean that these differences exist? What do the clusters analyses look like on the individual VAR networks? Many different clusters or homogeneous across

participants? [I assume "within one time point" refers to the sample-based estimate of the lag-1 autocorrelations? This is not "within one time point", but a time average, the contemporaneous network is the space average]

We thank the reviewer for this important comment, and we agree we did not interpret the results meaningfully. We think the reviewer makes an interesting point that we have investigated with additional analysis: we ran a cluster analysis on each person-specific VAR network model to see if the 3 assumed clusters would be found in the person-specific networks. We found heterogeneity amongst these results, i.e., for some individuals we found 3 clusters in their VAR model, for others 6, for some 18. The walktrap algorithm indicated a median cluster of 1 for person-specific temporal network structures (range: 1-18). For the contemporanous networks the walktrap algorithm indicated a median of 3 clusters (range: 1-18). For the spinglass algorithm we found a median of 6 clusters for the person-specific temporal network structures (range 2-10) and a median of 4 clusters for the person-specific contemporaneous network structures (range 2-6).

Given the extensive feedback of the editor and reviewers we have carefully reconsidered the aims of our paper as indicated in our preregistration and we have come to the conclusion that performing a cluster analysis does not contribute to the main aim "*Elucidate the working mechanism behind the interplay between affect and depressive complaints*" nor the subaims of our paper: (1) exploring the evolution of depressive complaints is accompanied by general trends in affect fluctuations and (2) to investigate the direct interplay between depressive complaints and affect in a more detailed level. Therefore, we have removed clustering analysis altogether from the manuscript. We do thank the reviewer for raising this point as it made us critically reflect on our aims and how the analyses we conducted contributed to these aims. We hope by investigating the reviewers concerns we have addressed them sufficiently. (See also our reply to reviewer 2 comment 7).
Regarding the note the reviewer makes within brackets, upon close reading we feel we may have referred to the contemporaneous network in a confusing manner. We agree that the term "within one time-point", that we used to indicate the contemporaneous network model might sound like the contemporaneous network is a lag(0) model. We have carefully reread the manuscript and have paid special attention to how we refer to the contemporaneous network and rephrased were necessary (e.g., see page 11):

"The mlVAR package estimates three types of network structures: (a) a temporal network (both a fixed-effect structure over all persons and a random-effect structure per person), (b) a contemporaneous network (both a fixed-effect structure over all persons and a random-effect structure per person), and (c) a between-persons network (Epskamp, van Borkulo, et al., 2018b). In line with Epskamp et al., (2018), for the remainder of this paper, the term within-person will refer to the fixed-effect within-

person network (either temporal or contemporaneous) and the term person-specific will refer to random-effects within-person networks (either temporal or contemporaneous).

3. In the recent literature the problem of integrating time series form different sources, or subsystems has become a topic of interest… this is often referred to as multi-layer or mulitplex networks. Perhaps the authors can comment in the discussion on whether the mlVAR method suffices to take of clustering or whether they should take other measures to deal with this problem, see e.g. Tio et al 2018 Introducing SNAC: Sparse Network and Component model for integration of multi-source data

Indeed, integrating time series from different sources or subsystems is challenging and may require more sophisticated statistical analyses, such as the SNAC method. The data that we aim to integrate, however, do not come from different sources. Both affect and depressive complaints are assessed using a single questionnaire. As such, we do not aim to estimate a multi-layer or multiplex network, and we also do not experience problems that may come from estimating these networks, such as very weak inter-layer correlations. Methods like SNAC have been developed precisely to overcome challenges that come from estimating networks from multiple sources, but since our data come from the same source, we do not see the added benefit of using SNAC. Moreover, currently SNAC cannot be used on longitudinal datasets and is therefore not suited for our current research question.

We do however agree that it is important to consider that they may operate on different timescales and we included this in discussion (see page 24):

"..affect and depressive complaints may operate on a different timescale. To the best of our knowledge, there are no current techniques available that can account for these possible time-scale differences in processes in the field of network analysis (Bringmann et al., 2021)."

4. If I understand correctly, the density was calculated based on individual networks for each participant…? If these networks were created and one of the objectives of the paper is to compare within versus between structures… why not report some of the well-known network measures as boxplots with the values of the mlVar networks as a reference?

We would like to thank the reviewer for pointing out that the objectives of the paper were not clearly described before. Our primary objectives are to "*to investigate how fluctuations in positive and negative affect are related to the evolution of depressive complaints during a prolonged period of stress"* (see page 6). In that context, we inspect the density of the person-specific networks (indeed, the individual networks for each participant) to "*investigate whether the structure of these person-specific networks*

is related to a clinically meaningful change (increase or decrease) of depressive complaints" (page 7). While interesting in its own, comparing within versus between network structures was not one of our objectives. We have now addressed this issue in the introduction where we, in line with our preregistration, describe the 3 aims of our paper and the related hypothesis (see page 6-7)

5. *"Interestingly, when considering the direction of change – alleviation or aggravation in PHQ-9 score – a bifurcation appeared, indicating that the same network density can relate to either a worsening or improvement in depressive complaints." * This is potentially an important finding that corroborates results from other studies that are not mentioned by the authors. It is related to the concept of the phase transition, which is the term I suggest should be used instead of "bifurcation". In the supplementary material Figure C1 (Figure 3. in the pdf has no points) does not show any "change" due to a control parameter leading to a bifurcation, it shows that the individuals with networks that are on average densely connected over the t>20 time points have, in the same time period, observed a large shift (either pos. or neg.).

We want to thank the reviewer for this valuable comment. Indeed, the term 'bifurcation' is not appropriate in the current context due to the lack of a control parameter. We agree that 'phase transition' would be more applicable. However, in the paper we focus on 'clinically meaningful change' and do not further discuss the different phases our model could have (from a more 'complex systems approach'). We therefore chose to remove the term bifurcation, and refer to meaningful change in an alleviating or aggravating direction instead of adopting another term (see page 20). We believe this puts the focus on our main result - that higher density correlates with both a clinically meaningful alleviation or aggravation of complaints - aligned with the more conceptual and clinical focus we aim to have in this paper. We outline how we incorporated this finding with the very valuable suggestions in the literature from the reviewer in the next comment (where we do discuss the possibility that the models show different *phase transitions*).

6. The increased connectivity or density can point to a reorganisation of the system in which previously inaccessible degrees of freedom are opened up (= destabilization) making it more likely the system can transition to a new stable state. This is why Early Warning Signals like an increase in invariance/critical fluctuations and an increase in the autocorrelation are associated with connectivity. This association was evidenced in a simulation study the authors cite (Cramer et al. 2016), but was empirically observed in a well-known N=1 study by Wichers & Groot(Critical Slowing Down as a Personalized Early Warning Signal for Depression 2016). In a N=300+ study by Olthof et al. (CriticalFluctuations as an Early-Warning Signal for Sudden Gains and Losses inPatients Receiving Psychotherapy for Mood Disorders ClinicalPsychological Science 2019) it was shown that the presence of EWS predicted sudden

gains as well as sudden losses in patients with mood disorders. I think the impact of the paper would be increased if the authors can connect their results to these (and other) findings that are already present in the empirical record.

We are very grateful for the valuable suggestions by the reviewer and would like to thank them for directing us towards the literature. We agree with the reviewer that this explanation was currently lacking in the discussion and incorporated this explanation in the discussion. See (page 23):

"The strong relation between network density and change in depressive complaints may have important implications for the clinical interpretation of networks, as network density has generally been related to more severe psychopathology (e.g., see Calugi et al., 2021; van Borkulo et al., 2015). However, our study shows an alternative situation in which a larger density of networks indicates more fluctuations and potential for flexibility (Hayes et al., 2015). One possible explanation for our finding is that phase transitions in a wide variety of systems (e.g., transitioning from mild depressive complaints to severe complaints) are often characterized by a period of instability, in which the behavior of system shows many fluctuations (Olthof et al., 2020; Wichers et al., 2016). This increase in fluctuations before a clinically meaningful alleviation or aggravation could be reflected in the increased network density. This explanation possibly corroborates findings in other studies that found larger densities in networks to be related with a decrease in psychopathology symptoms over time (McElroy et al., 2019)."

Additionally, our methodological explanation on test sum scores and covariations is nuanced and presented as another alternative (see page 23):

"While the relation between network density and change in depressive complaints may signal relevant clinical importance, it is important to note that this strong relation could also merely reflect a well-known property of test reliability, namely that the variance of a total score (in our case the change in depressive complaints) consists of the sum over the variance in all items (in our case the individual affects and individual PHQ-9 items) and the sum over their covariances (Cronbach, 1951). Since denser networks reflect stronger covariances, it is a statistical necessity that denser networks are accompanied by larger variations in the total score (i.e., the variation in PHQ-9 score). However, our finding does not merely reflect a methodological artefact, as our sensitivity analysis indicate that only including the affect states in the network structure (thus, removing the PHQ-9 items from the network) still rendered a positive association between network density and absolute change in PHQ-9 score (see Supplement C for more details on the performed sensitivity checks)."

Reviewer 2:

1. Outline their rationale more clearly in the introduction (e.g., what questions does this work aim to answer), provide an extended discussion of their current work as it is informed by previous work in the field (e.g., how is their study connected to previous network analyses on affect and depression) and discuss the implications of this work (e.g., how will their findings contribute to the field)

We agree with the reviewer that the description of our theoretical foundation in the introduction was not solid enough, and we want to thank the reviewer for pointing this out. We have greatly re-written the introduction, and will highlight specific sections here that reflect the requested changes above: (i) outlining of our rationale, (ii) providing an extended discussion, and (iii) discussing the implications of our work.

We outlined our rationale in the introduction by first adding a substantive part on the relationship between affect regulation and the development of depressive complaints in the literature (addressing point ii on providing an extended discussion; page 3-4). We also explained better why we use network analysis, as it allows us to look into various direct and indirect pathways of how fluctuations in affect relate to the evolution of depressive complaints see (pages 4-5 for a more detailed explanation). Following these extended discussion on the relation between affect and depressive complaints as well as our rationale for using network analysis, we explicitly state our research aims (addressing point i on outlining our rationale):

Overall aim (page 4): "*We aim to deconstruct the relationship between affect and depressive complaints in the face of stressful times and take a step further in understanding the heterogeneity of depression."*

Sub-aims:

"We first aim to explore whether the evolution of depressive complaints in the beginning of the COVID-19 pandemic (i.e., worsening or alleviation of complaints) is accompanied by general trends in affect fluctuations over that same period." (page 6)

"Second, to investigate the direct interplay between depressive complaints and affect over the studied period at a more detailed level, we model their relations more directly using longitudinal multilevel network models." (page 6)

"..., we aim to investigate whether the structure of these person-specific networks is related to a clinically meaningful change (increase or decrease) of depressive complaints." (page 7)

Finally, in the discussion we discuss further implications of our work and how our findings will contribute to the field, especially regarding our finding that network density is related to overall change in depression complaints (see page 23):

"The strong relation between network density and change in depressive complaints may have important implications for the clinical interpretation of networks, as network density has generally been related with more severe psychopathology (e.g., see Calugi et al., 2021; van Borkulo et al., 2015). However, our study shows an alternative situation in which a larger density of networks indicates more fluctuations and potential for flexibility (Hayes et al., 2015)."

2. Clearly link their research questions to stated hypotheses (per their pre-registration) and how/why their selected methods can test their hypotheses.

We agree with the reviewer that the introduction needed more clarity on our aims, hypotheses and methods. We now clearly state our aims in the introduction, explain in more detail how each sub-aim will be investigated, what methods will be used, and what our hypotheses are (per our pre-registration) (see page 6-7). In addition, in the Discussion section, we reflect back on our (rejected) hypotheses (see page 21-22):

"As expected, we found differences in the affect trajectories among people who experienced consistently no depressive complaints (showing higher PA than NA) compared with consistent depressive complaints (showing the reversed pattern). Crucially, these differences pertained to both PA and NA trajectories, showing that there is a clear link between consistent severity levels of depressive complaints and both positive and negative affect. Contrary to our expectations, affect trajectories were similar for people experiencing either an aggravation or alleviation of depressive complaints, meaning that we reject our hypothesis that people who experienced an increase in depressive complaints would show high levels of NA, whereas people who reported a decline in complaints would show high levels of PA."

3. I was often confused by the use of the label "individual" to describe findings that appears to within-person. Multilevel vector autoregression measures network associations between and within individuals over time, but do not assess network structure for specific individuals (Epskamp et al.,2018). I recommend the authors rephrase "individual" to "within-person" for consistency with standard notation for multilevel modeling approaches and to increase the clarity of their aims and findings.

We are sorry for the confusion we have caused with using the label individual to describe findings that appear on the within-person level. We now clearly distinguish between the different types of networks the ML VAR model estimates. There are 6 network structures that can be obtained when estimating a ML VAR mode: a (1) between-person level network, (2) average within-person contemporaneous network, (3) average within-person temporal network, (4) person-specific contemporaneous networks, (5) person-specific temporal networks, and, although not relevant for the current paper but worth mentioning for completion, a (6) standard deviation of random-effects network. We now describe the types of network structures that are estimated with the mlVAR in more detail in the methods section. In line with Epskamp et al., (2018), the term within-person now always refers to the (average) fixed-effect within-person level network structure (either contemporaneous or temporal) while the term person-specific refers to random-effect within-person level network structure (either contemporaneous or temporal) that is estimated per person using ml VAR. We explain this in our method section as follows (see page 11):

"The mlVAR package estimates three types of network structures: (a) a temporal network (both a fixed-effect structure over all persons and a random-effect structure per person), (b) a contemporaneous network (both a fixed-effect structure over all persons and a random-effect structure per person), and (c) a between-persons network (Epskamp, van Borkulo, et al., 2018b).). In line with Epskamp et al., (2018), for the remainder of this paper, the term within-person will refer to the fixed-effect within-person network (either temporal or contemporaneous) and the term person-specific will refer to random-effects within-person networks (either temporal or contemporaneous)."

4. Measures: It appears that affect was assessed in the current moment; however, depressive symptoms were assessed over the last several days. What is the impact on interpretability when the prompt "over the last several days" is being used to assess items (near) daily?

We agree the reviewer raises an important difficulty here that was previously not sufficiently addressed in our paper. Although both depressive complaints and affect were assessed at the same time using the same questionnaire, questions regarding affect were prompted: "indicate which description best describes how you currently feel, right now in the moment" while questions regarding depressive complaints were prompted: "In the last several days, how often have you been bothered by any of the following problems". We now clearly describe this in the result section and provide the impact of this on the interpretation of the relations we found in the network (see page 17-18):

"For the interpretation of the relations within each network structure, it is important to note that questions regarding affect and depressive complaints were prompted differently. Affect was prompted: "Indicate which description best describes how you currently feel, right now in the moment" while

questions regarding depressive complaints were prompted: "In the last several days, how often have you been bothered by any of the following problems". This difference in phrasing, as well as the temporal scaling of our study (roughly 3 days between assessments) influences the interpretation of the estimated networks. The relations in the average contemporaneous network reflect how depressive complaints over the days prior to the assessment and affect at the moment of assessment co-occur (e.g., how depressive complaints over a period of 3 days, let's say Monday through Wednesday, co-occurs with affect on Wednesday). Relations in the average temporal network reflect how depressive complaints over the days prior to the assessment predicts affect at the next assessment occasion (e.g., how depressive complaints over a period of 3 days, for example, Monday through Wednesday, predicts affect on Friday) and vice versa. The relations in the between-person network reflect how, on average, depressive complaints relate to affect."

We refer back to the differences in phrasing for measuring affect and depression in our discussion (see page 24):

"*Third, affect and depressive complaints may operate on different timescales. To the best of our knowledge, there are no current techniques available that can account for these possible time-scale differences in processes in the field of network analysis (Bringmann et al., 2021). However, the time-scale differences in affect and depressive complaints where taking into account in the questionnaire. Questions regarding affect were prompted: "Indicate which description best describes how you currently feel, right now in the moment" while questions regarding depressive complaints were prompted: "In the last several days, how often have you been bothered by any of the following problems". This difference in phrasing captures the idea that affect operates on a faster time-scale (e.g., fluctuates within a day), while depression operates on a slower time-scale (e.g., fluctuates from day-to-day).*"

5. Rationale for subgrouping/Berkson's bias: Subgrouping based on depression severity would seem to increase the risk of Berkson's bias, which can result in spurious findings (e.g., see De Ron et al., 2019). It seems that the authors' aims could be achieved equally well by testing a continuous moderating effect of depression symptom severity on affect using mixed graphical modeling as outlined in Aim 3 of their pre-registration.

For the estimation of the ML VAR network we have not grouped our participants based on depression severity, therefore we do not run into the risk of inducing Berksons bias. We realise this is a deviation from the analysis plan as described in our pre-registation. We deviated from this analysis plan for the exact reason the reviewer now highlights. Instead, we inspect the trajectories of affect and depression for different courses of depressive complaints using Loess regressions.

In addition, we apologize for the confusion we have caused regarding our aims. We hope to have taken away this confusion by clearly stating our aims, our hypothesis and linking these aims to the analysis in the current paper as we described to the reviewer in comment 2 (see page 6-7 in the manuscript).

The third aim ("To study the potential effects of risk and protective factors on fluctuation levels of depression) from our pre-registration, is not included in the current paper. To investigate this aim, we wanted to include risk and protective variables such as substance abuse, sleep problems, emotion regulation, and exercise, to a symptom network of depressive complaints using the MGM network estimation technique. We have decided not to cover this aim in the current paper as it required the introduction of new variables (e.g., the protective and risk factors), new literature (e.g., resilience literature) and new estimation techniques (e.g., MGM and moderation effects). Instead, we are working on aim 3 in a separate paper.

6. Composition of the selected subsample: Per my understanding of their sample selection process, it appears that only 17% of the sample(approximate) was retained for the main analyses. It would be helpful for the authors to provide some indication of how the subsample differed in demographic or clinical characteristics to the larger sample o It would also be helpful for the reader to review the subsample's mean depression symptom severity at baseline (e.g., in Table 1)

We agree with the reviewer (and editor) that it is important to report whether and how the included participants differed from those excluded. As replied to the editor, we have now included sample characterisitcs (see Table 1; page 14-15, and Table B1; page 39-40), and means and standard deviations of the variables of interests (see Table 2; page 15 and Table B2; page 40) for both the original sample and the selected sample. For our original sample these tables have been added to a supplement, for the selected sample the tables have been added to the manuscript. In addition we highlight (see page 13-14):

"Our sample (n=228) differed from the original full sample (n=1355) in their mean PHQ-9 score at baseline; the original full sample had a higher PHQ-9 mean at baseline. Our sample did not differ in their PANAS scores at baseline from the original full sample (n=1355), see Supplement B, Table B1 for more details on sample characterization, means standard deviations, and range for the PHQ-9 and PANAS for the original full sample."

7. Multicollinearity: Could the authors present (or refer the reader to) a zero-order correlation matrix demonstrating the associations between items included in the network? It was not clear whether multicollinearity between items (r > .6) may have impacted their findings (see Fried &

Cramer, 2017). If so, it may be necessary for the authors to composite items with large zero-order correlations using a node-reduction method (e.g., goldbricker, Jones,2020). Notably, the authors used the Spinglass and Walktrap clustering algorithms in a series of sensitivity analyses (page 26). I had some difficulty following the rationale for these sensitivity analyses (i.e., what specific question were these analyses intended to answer; was the purpose to address multicollinearity?) I encourage the authors to clarify their rationale for these analyses in both the main and supplementary text. Further, there are criticism of the Spinglass algorithm in psychological data (Christensen etal., 2020), in part as it assumes only one community per item (Fried, 2016). It would be helpful for the authors to explain whether and how the criticisms impact their findings.

We would like to thank the reviewer for these considerations. As requested, we inspected the zero-order correlation matrix for the baseline assessments of the variables included in our network. For some of our items we found a correlation higher than 0.6, however, following recommendations (Borsboom et al., 2022), including certain variables in our network is a choice primary driven by substantive rather than methodological considerations. We included the items from the validated short-form PANAS and are, therefore, hesitant to remove certain items post-hoc. Moreover, high correlations in the network are not impacting any of our conclusions, as could potentially be the case if we would compute 'strength centrality' for example: strength centrality would be inflated for items that potentially show topological overlap (Fried, 2016). We hope the reviewer agrees with our careful considerations to not remove items from the network.

We fully agree with the reviewer that the rationale for sensitivity analyses like the Walktrap or Spinglass community detection was insufficient. Given the extensive feedback of the editor and reviewers (see also our reply to reviewer 1, comment 2) we have carefully reconsidered the aims of our paper as indicated in our preregistration and we have come to the conclusion that performing a community detection analysis does not contribute to the main aim "*Elucidate the working mechanism behind the interplay between affect and depressive complaints*" nor the subaims of our paper: (1) exploring the evolution of depressive complaints is accompanied by general trends in affect fluctuations and (2) to investigate the direct interplay between depressive complaints and affect in a more detailed level. Therefore, we have removed the community detection analysis altogether from the manuscript. We do thank the reviewer for raising this point as it made us critically reflect on our aims and how the analyses we conducted contributed to these aims.

8. Stability/reliability of network analyses: It would be helpful for the authors to test and present the results from reliability analyses that demonstrate the stability of their networks within the subsample. One suggested process is described here (Jongeneel et al., 2020)

We thank the reviewer for their suggestion and have examined the stability of our results following a similar procedure as described in Jongeel et al., (2020). We describe the procedure we followed to assess the stability in the method section, see page 13:

"To examine the stability of the correlation between person-specific temporal network density and change in PHQ-9 score we followed a similar procedure as described in Jongeneel et al., (2020). We re-estimated 100 mlVAR networks that included a random selection of 80% of the original data (i.e., using the data of ~182 participants). For each of these re-estimated networks we followed the same procedure as described above, (i.e., we calculated person-specific network density and correlated the network density to change in PHQ-9 score)."

In addition, results for the stability analysis are discussed in the results section, see page 20-21:

"Stability checks indicated the originally found correlation between person-specific network density and maximum change in PHQ-9 score was stable. The strong correlation between absolute maximum change in PHQ-9 and persons-specific temporal network density was still present when recursively re-estimating networks using 80% of the data at a time. The median correlation between person-specific network density and their maximum change in PHQ-9 score using 80% of the data was R = 0.77 and ranged from 0.72-0.81. In addition, we found stable results for the relation between person-specific network density and the aggravation and alleviation of depressive complaints; when looking at the correlation between maximum change in PHQ-9 and person-specific network density, we found a median correlation of R = -.04, ranging from-0.15 to 0.07. (See Figure C3 in Supplement C for the correlation plots)."

We would like to emphasise again that we have not estimated networks in subsamples as previously indicated in our pre-registration, see our response to comment 5.

9. Calculation of individual-level temporal density: Could the authors clarify how they calculated "individual density" (page 9)? It appears that they constructed temporal networks for each individual and then calculated the network strength from the individual-level network. If that is the case, then I am concerned that the findings from these temporal networks are critically under powered if the average number of individual-level observations was approximately 20 (Mansueto et al., 2022). Without a sufficient number of observations to construct stable individual-level networks, the strength and accuracy of their inferences regarding these temporal networks appears critically limited.

We agree with the reviewer that the estimation of pure idiographic networks using a graphical VAR method requires more observations per individual than 20 as is shown in Mansueto et al., (2022). Therefore, we did not estimate pure idiographic network structures using the graphical VAR method. Instead, we estimated person-specific networks using a multilevel VAR estimation technique. This estimation technique requires less observations per individual (The method performs well from 20 observations per individual onwards (Epskamp et al., 2018)) as information is borrowed from other participations in order to estimate networks at both the between-level as well as on the within-level (as fixed-effects, indicating average effects and random-effects, indicating person-specific effects). We used these random-effect network models to calculate the person-specific network density.

We now realise the term "individual network" might be confusing and therefore refer to the random-effect within-level network structures as person-specific network structures. We have clarified the terms used in our paper in several ways, as mentioned in our reply to comment 3 (see page 11):

"In line with Epskamp et al., (2018), for the remainder of this paper, the term within-person will refer to the fixed-effect within-person network (either temporal or contemporaneous) and the term person-specific will refer to random-effects within-person networks (either temporal or contemporaneous)."

Furthermore, person-specific network structures as estimated using the ml VAR methods, should not be taken as an indication of pure idiographic networks. We emphasise this point in the discussion (see page 24-25):

"In addition, it should be noted that although multilevel VAR estimation allows for the estimation of person-specific networks, these networks are not purely idiographic. With multilevel estimation, a shared underlying distribution for all model parameters is assumed. This means person-specific edge weights are "shrinked" to follow the same underlying group distribution. In turn this means, individual differences are smoothed out in the estimation process. Therefore, future research should indicate whether the results found here hold on a purely idiographic level."

10. Deviations from pre-registration: Per the pre-registration, it appears that the authors described a plan to use mgm to test moderation effects betweenthe risk and protective factors and the symptom network. However, in my review of the manuscript, I did not see any description of their use of MGM.Could the authors clarify whether this was a deviation from the pre-registration and if so, share their decision process?

We highly appreciate the reviewer for taking on the effort of comparing our pre-registration with our manuscript. The reviewer is right, the manuscript submitted here deviates from our analysis plan as

described in the pre-registration (aim 3, linking our results to risk and protective factors). During the process we found that our first two aims, focusing on affect and depressive complaints, were already quite expansive. We were afraid that adding more variables to the mix as potential risk and protective factors would complicate matters and not add to the readability of the paper, as the theoretical foundation for the choice of these variables should also be included in the introduction and aims. We decided that aim 3, including other variables of the questionnaire as potential risk and protective factors, would better fit in a different paper on which we are currently working (see also our reply to comment 5 and comment 8 in this letter).

11. Tests of stationarity: I appreciated that the authors tested for trends in their data (as presented in the supplementary data). Could the authors state whether the kpss.test revealed significant trends in the data (page 29)? Additionally, given that stationarity is a key assumption of network analytic approaches, I encourage the authors to present these findings in the main text (as part of the data analytic plan or results) to support their use of network analyses to examine associations in these data.

We have stated whether the kpss.test revealed significant trends in the data (see Supplement Cl, page 41):

"As a first sensitivity check, we inspected whether trends in the data influence the observed correlation. To investigate this, we first tested for trends in the data using kpss.test() in R. The kpss.test revealed significant trends in the data."

In addition, we added a subheading to the methods section describing the sensitivity analyses performed and the rationale behind them (see page 12-13):

"To examine the extent to which the observed correlation between person-specific temporal network density and their change in PHQ-9 score could be dependent on methodological choices we conducted two sensitivity checks. First, we determined whether potential trends in the data could have affected our observed correlation. We checked for trends in the variables and detrended any variables with significant trends. Then we followed the same procedures and computed the correlation. Second, since the network includes the PHQ-9 items we examined if the correlation between network density and maximum change in PHQ-9 score might be driven by this overlap. Given this overlap, the temporal density is in part based on the same information (i.e., the PHQ-9 items) as the change score in PHQ-9 items. To that end, we re-estimated a mlVAR network including only the affect states as nodes and correlated the person-specific temporal network density of the affect states with participants' change in PHQ-9 scores."

Finally, we now present the results of our sensitivity checks in our main findings (see page 20):

"Sensitivity checks showed the strong correlation between absolute change in PHQ-9 and person-specific temporal network density was still present when data was detrended (R = 0.79). When only affect states were included in the network structure, the strength of the observed correlation decreased (R = 0.4), however we still observed a positive association between person-specific network density and absolute maximum change in PHQ-9 score.

In addition, the relation between network density to both a more substantial aggravation and a more substantial alleviation in depressive complaints was still present when data was detrended, or when only affect states were included in the network structure. See Supplement C for more details on the performed sensitivity checks."