Review of the Guidance Document for the risk assessment for bees

Supporting document for Risk Managers consultation on Specific Protection Goals for bees

Analysis of background variability of honey bee colony size

17 December 2020

Preliminary report

The analysis presented in this preliminary report was carried out by the experts of the EFSA Working Group for the revision of the bee guidance (i.e. EFSA staff, bee and exposure experts and regulators) with the external support of a PPR Panel member with expertise in modelling.

This preliminary report is provided to risk managers for the next step of the consultation on the specific protection goals to progress with the review of the EFSA (2013). A full analysis will be reported and published in a dedicated technical report.

European Food Safety Authority
Via Carlo Magno 1A – 43126 Parma, Italy
Tel. +39 0521 036 200| www.efsa.europa.eu

Contents

Summary **3**

1. Background **5**

2. Scope of the document **5**

3. General framework for the review of the SPG for honey bees **6**

3.1. Defining SPGs based on the EFSA method 6

3.2. Implementation of the SPG in the risk assessment 8

3.2.1. Reference tier 8

3.2.2. Tiered approach and trigger values 9

3.3. SPG dimensions with Approach #2 10

3.3.1. Informing the definition of the "magnitude" dimension using the concept of operating range (OR) 11

4. Materials and methods **13**

4.1. The use of the BEEHAVE model 14

4.2. Environmental scenarios 15

4.3. Description of the scenarios 15

4.4. External data 17

4.4.1. Data used for scenario calibration 17

4.4.2. Data used for checking the plausibility of the model simulations 18

5. Results **18**

5.1. Overall results of the simulations 19

5.2. Interpretation of the results 20

5.2.1. Recommendations on how to interpret the results 21

5.3. Plausibility of the model simulations 22

6. Uncertainties and potential future developments **23**

6.1. Limitations of BEEHAVE identified in EFSA PPR Panel (2015) 24

6.2. Limitations of BEEHAVE identified in the present analysis 24

6.3. Relevance of input values for the present analysis 25

6.4. Uncertainties in the scenario definition 25

6.5. Outlook 27

7. Reference tier (field studies) design in relation to the magnitude of acceptable effect **28**

7.1. Preliminary estimation of the requirement for higher tier studies 28

7.2. Example from available higher tier studies 28

7.3. Considerations of the requirements of field studies in the EFSA bee guidance document 29

8. Concluding remarks for decision making process for risk managers **29**

References **32**

Appendix A – Results of the simulation per regulatory zone **35**

Appendix B – Analysis of landscape complexity **38**

Appendix C – Variability in risk assessment **43**

Summary

Risk managers agreed that background variability in colony size can be used for defining Specific Protection Goals for honey bees

This document describes a method for defining Specific Protection Goals (SPGs) for honey bees by deriving the SPGs from the background variability of colony sizes. It allows risk managers to set SPGs which contain the impact of pesticides on the number of bees within the range of the background variability of the colony sizes.

In the context of the definition of SPGs for bees, risk managers asked EFSA to provide scientific background to support them in their decision-making process about what needs to be protected and to what extent. Among the four approaches that EFSA developed, the risk managers indicated that the derivation of a threshold of acceptable effects on colony size based on their variability (i.e. approach #2) was the preferred option for honey bees. This approach assumes that when evaluating a pesticide, the magnitude of acceptable effects should be set within the range of the background variability of colonies not exposed to pesticides. In this way, it is assumed that any impact on the provision of the ecosystem services depending on honey bees would also remain within the background variability.

EFSA used BEEHAVE to assess background variability of colony size in multiple scenarios

The analysis was performed with the BEEHAVE computer model in nineteen EU environmental scenarios covering a range of geographical, climatic, and beekeeping conditions. For each scenario, 500 replicate simulations were run under equal conditions. Each replicate showed the dynamics of a honey bee colony over one year, in situations where the bees were not exposed to any pesticide. The outcome of the simulations for each scenario is conceptually comparable to the observations of replicate hives in the control group of experiments field studies, which are the reference for the risk assessment for bees.

The background variability allows risk managers to set the level of protection for colonies exposed to pesticides

Plotting the modelled development of the colony sizes over time gives a picture of their background variability. From the variability distribution two elements are of interest: the mean colony size and the lower end of the distribution. The difference between these two values defines the extent to which the size of a colony can be reduced because of background variation. In practice, the knowledge of the shape of the distribution curve allows limits to be set for the reduction in colony size caused by exposure to pesticides. Variations within the limited range would be considered as acceptable.

The results of the analysis are presented for the whole year as well as for each season and for each regulatory zone. A summary for the scenarios with the minimum and maximum variability is also presented, leading to percentage ranges that can inform on colony size reduction. These percentages were calculated for the full variability distribution and for several restricted variability distributions. With a more restricted variability range, the threshold of acceptable effects is more conservative. The analysis of the simulated background variability distribution shows that a large fraction of the total variability is caused by a limited number of colonies.

Additional elements may support the decision of the risk managers: uncertainties, comparison with experimental data, and practical implementation in field studies

The model combines fixed input parameters and stochastic elements. For some elements belonging to both categories, uncertainties were identified, but could not be fully evaluated and quantified within the present work. However, a comparison was made between the model outcome and the measurements performed on control groups of experimental field studies. Such comparison shows that the model predictions were in the range of the experimental values, but there was a general underestimation of the median variability. A further analysis showed that the variability increases with increasing landscape complexity. The simulated nineteen scenarios are all characterised by a very simple land-

scape, thus confirming that variability is likely underestimated compared to the real world. In the present context, an underestimation of the variability leads to a more conservative threshold of acceptable effects.

The analysis of the background variability presented in this document should support risk managers in defining a threshold equivalent to a certain percentage reduction in colony size that is considered acceptable, in a similar way as was proposed by EFSA (2013). This threshold represents the largest acceptable mean colony size reduction that exposed colonies can suffer when compared to the unexposed colonies in the control. The threshold will be used to evaluate the field studies; therefore, it should be implementable and measurable. The selected threshold of acceptable effects will determine the requirements for the design of field studies. This document makes explicit the link between the threshold and the complexity of the study design, along with a benchmarking of recent state-of-the-art field studies described in the scientific literature.

1. Background

In the context of the definition of Specific Protection Goals (SPGs) for bees, risk managers asked EFSA to provide scientific background to support them in their decision-making about what needs to be protected and to which extent. In the first supporting document[1], published at the end of July 2020, EFSA described a set of four possible approaches developed to respond to the request.

The four approaches are possible scientific processes which risk managers could choose to determine SPGs. These approaches were developed by considering the request of the European Commission mandate to *take into account planned and ongoing discussions initiated by the Commission on defining specific environmental protection goals and review the risk assessment guidance based on the specific protection goals agreed during this process* (ToR6).

EFSA took into consideration the preliminary outcome of the action initiated in 2019 by the European Commission towards defining SPGs involving member states (MSs) and stakeholders; in particular, the positive opinion conveyed by stakeholders and MSs on the use of the EFSA framework for identifying SPGs[2]. Based on this preliminary outcome, EFSA deemed that a full review of the SPG defined in the EFSA bee guidance document[3], involving all steps of the EFSA method[2], may not be necessary and was considered outside of the scope of this mandate. In fact, the EFSA method[2] for defining SPGs was already implemented in the EFSA bee guidance document[3]. Nevertheless, EFSA has elaborated the four approaches, along the lines of this preliminary outcome, to address the feedback from MSs on the SPGs as defined in the EFSA bee guidance document[3] and to support the risk managers on the revision of some of the five dimensions, i.e. Step 3 of the EFSA method[2].

The four approaches were presented on 30 June 2020 to the representatives of the MSs in a workshop organised by the European Commission. They are summarised below:

- Approach 1 – to establish acceptable effect based on long-term colony survival;
- Approach 2 – to derive a threshold of acceptable effect on colony size based on background variability;
- Approach 3 – to establish acceptable effect, based on predefined levels, on colony/population size;
- Approach 4 – to establish acceptable effect on colony/population size based on levels of acceptable impact on the provision of ecosystem services.

The scientific concepts underlying each approach, reported in the first supporting document[1], were explained to risk managers and discussed during the workshop in June 2020. The pros and cons were also described along with the analysis of the feasibility of their implementation within the timeline of the current mandate.

As a result of the discussion, a large majority of the MSs expressed a preference for approach #2 for honey bees. This choice was confirmed at the meeting of the Standing Committee on Plants, Animals, Food and Feed, Section Phytopharmaceuticals – Legislation (SCoPAFF) on 16 July 2020.

EFSA presented the four approaches to the stakeholder ad-hoc group in an information session organised on the 23 September 2020.

2. Scope of the document

In the present document, approach #2 and its implementation are presented. It is important to bear in mind that the outcome of the implementation of approach #2 presented in this report focuses on honey bees, and cannot be used for defining SPGs for bumble bees and solitary bees, due to their different biology and ecology e.g. smaller colony size for bumble bees, solitary nesting in contrast to

[1] https://www.efsa.europa.eu/sites/default/files/topic/EFSA-Supporting-document-for-RMs-in-defining-SPGs.pdf
[2] EFSA Scientific Committee (2016) and EFSA PPR Panel (2010)
[3] EFSA (2013)

colony formation for solitary bees, shorter nesting periods, feeding and breeding behaviour (EFSA PPR Panel, 2012).

The implementation of the principles of approach #2 for bumble bees might be considered at a later stage after suitable models e.g. the Bumble-BEEHAVE model (Becher et al. 2018) have been evaluated according to the EFSA good modelling practices opinion (EFSA PPR Panel, 2014).

On the basis of current knowledge, approach #2 cannot be used for the bumble bee and solitary bee groups. As explained in the first supporting document[1], due to the lack of knowledge and data, EFSA cannot provide further scientific grounds in this document for supporting the risk managers' decision on SPGs for bumble bees and solitary bees. Therefore, in the context of the review of the SPGs, risk managers could consider adopting a pragmatic approach for solitary bees and for bumble bees. EFSA PPR Panel (2012) suggested the application of uncertainty factors to the effect percentages identified for honey bees as a pragmatic solution.

Summary box 1
Non-*Apis* bees The analysis presented in this document is not suitable for defining the SPG for bumble bees and solitary bees.

3. General framework for the review of the SPG for honey bees

3.1. Defining SPGs based on the EFSA method

In 2019, the Commission initiated actions towards defining SPGs involving MSs and stakeholders on the basis of the EFSA method[2] for defining SPGs. The EFSA method outlined includes several steps:

Step 1 – identification of the relevant Ecosystem Services (ES) potentially impaired by a stressor;

Step 2 – identification of the relevant Service Providing Units (SPU);

Step 3 – specification of the level/parameters of protection of the SPUs based on five interrelated dimensions: 1) Ecological entity; 2) Attribute; 3) Magnitude of the effect; 4) Temporal scale; 5) Spatial scale.

So far, the Commission has organised three workshops: two in 2019 with MSs and stakeholders separately, and one in February 2020, with both stakeholders and MSs. Generally, stakeholders and MSs were positive about the use of the EFSA framework[2] for identifying SPGs. In the workshop in February 2020 step 1 of the EFSA method was discussed for different pesticide use scenarios. The provision of pollination was widely recognised as a key ecosystem service.

Already in the EFSA bee guidance document[3] and in the preceding Scientific Opinion[4], ecosystem services and SPGs were identified and discussed with risk managers, according to the methodology proposed by the EFSA opinion for SPGs[5]. Therefore, the methodology and the process implemented in the EFSA bee guidance document[3] can be considered in line with the EFSA method[2] for defining SPGs and therefore with the action initiated by the European Commission. The ecosystem services identified for the EFSA bee guidance document[3] were **pollination**, **food and genetic resources provisioning, and cultural services.** These are in line with step 1 of the EFSA method[2] as discussed at the workshop held with stakeholders and MSs in February 2020.

Furthermore, the EFSA bee guidance document[3] includes, beyond honey bees covered in the current data requirements[6], bumble bees and solitary bees. This means that the second step of the EFSA method[2] – identifying the SPU for the above ecosystem services – can already be considered as partially addressed. As a general remark, additional SPU may be added, i.e. other pollinators that are not covered by the EFSA bee guidance document[3] if identified as being important to be covered by future guidance.

[4] EFSA PPR Panel (2012)
[5] EFSA PPR Panel (2010)
[6] Regulation 283/2013 and 284/2013

The EFSA opinion[4] suggested a specification of five interrelated dimensions of the SPG (i.e. *Ecological Entities, Attribute*, *Magnitude*, *Temporal and Spatial scale*) in line with the third step of the EFSA method[2] (see Table 1 for details). These were discussed with risk managers and implemented in the EFSA bee guidance document[3]. Some of these dimensions may need to be discussed again by risk managers.

Table 1. Overview of the SPGs as implemented in the EFSA bee guidance document (EFSA, 2013) and defined in the preceding scientific opinion (EFSA PPR Panel, 2012) in light of the steps described in the EFSA framework for defining SPGs (EFSA Scientific Committee, 2016).

<table>
<tr><th>EFSA Scientific Committee (2016)</th><th>EFSA PPR Panel (2012)
EFSA (2013)</th></tr>
<tr><td>Step 1
Definition of ecosystem services</td><td>Pollination, food and genetic resources provisioning, and cultural service.</td></tr>
<tr><td>Step2
SPU</td><td>Honey bees, bumble bees and solitary bees</td></tr>
<tr><td>Step3
Specification of the level/parameters of protection of the SPUs based on five interrelated dimensions</td><td>
<table>
<tr><th><u>Dimensions</u></th><th><u>Honey bees</u></th><th><u>Bumble bees</u></th><th><u>Solitary bees</u></th></tr>
<tr><td>Ecological Entities</td><td><u>colony</u></td><td><u>colony</u></td><td><u>population</u></td></tr>
<tr><td>Attribute</td><td>Colony strength*</td><td>Colony strength*</td><td>Population abundance</td></tr>
<tr><td>Magnitude</td><td>Negligible – effect**</td><td>Negligible – effect**</td><td>Negligible – effect**</td></tr>
<tr><td>Temporal scale***</td><td><u>Not relevant</u> i.e. any time</td><td><u>Not relevant</u> i.e. any time</td><td><u>Not relevant</u> i.e. any time</td></tr>
<tr><td>Spatial scale</td><td>edge of field</td><td>edge of field</td><td>edge of field</td></tr>
</table>
*Colony strength is defined operationally as the number of adult bees in a colony (= colony size).

**negligible in the EFSA (2013) is such if statistically distinguishable from "small effects". The effect was considered negligible when the magnitude is below 7%.

it is important to note that the above SPGs and in particular, the Magnitude of the effect (i.e. effect sizes) have been defined principally by reference to honey bees. In the case of other bees, the same magnitude has been used as a surrogate to colony-level impacts (for other social bees, such as bumble bees) or to population abundance (solitary bees).

*** Based on EFSA PPR Panel (2010) and EFSA Scientific Committe (2016), no temporal scale is relevant if the selected magnitude is "negligible"
</td></tr>
</table>

3.2. Implementation of the SPG in the risk assessment

3.2.1. Reference tier

The EFSA method[2] for deriving the SPGs, and in particular EFSA PPR Panel Opinion[5], suggests identifying for each SPU a reference tier for developing the risk assessment scheme. The reference tier is represented by the most sophisticated experimental or modelling risk assessment method that addresses the specific protection goal, and is then used to calibrate lower tiers which are based on simpler methods that are practical for routine use. In a routine risk assessment, the reference tier would only be used when the results of the lower tiers do not demonstrate a low risk for a specific use.

In the case of honey bees, **the reference tier is represented by field studies**. These are experiments with a high level of realism, characterised by complex set-up and interpretation of the results. In general, these studies aim to compare at least two groups of honey bee colonies:

a) The treated group, which is exposed to the pesticide under investigation. Field studies are performed with the aim of mimicking realistic conditions. This implies that the pesticide under investigation is applied to a crop that the honey bees have access to for collecting pollen and/or nectar. The pesticide application rate, frequency and timing should be in line with the use for which authorisation is sought.

b) The control group, which is set up in the same way as the treated group with the exception that it is not exposed to the pesticide under investigation. The control group should have access to the same cropping system / field characteristics as the treatment group, but these do not receive the treatment of the pesticide being investigated.

While more complex designs are possible (e.g. combining investigations in several regions at the same time, etc.) the underlying basic principle remains a comparison between the treatment and the control group.

The reference tier should be able to address the defined SPG – in all its dimensions – by performing targeted measurement. All five dimensions of the SPG contribute significantly to the design of field studies.

The **ecological entity** identifies the object of the experimental observation. For honey bees, these are the colonies.

The **attribute** identifies the main variable to be measured. This is not necessarily the only measured variable, but it is the one driving the overall risk assessment. In the EFSA bee guidance document[3], this was colony strength, which was operationally defined as the number of adult bees in a colony (= colony size).

The **magnitude** of the effect is pivotal both in the design of the study and in the interpretation of the results. The most straightforward way to check whether the exposure to a certain pesticide caused an effect on the colony strength is to compare the arithmetic mean value of this variable in the control and in the treatment groups. A difference larger than the agreed magnitude is an indication that the SPG may not be met in the study. Furthermore, the definition of a certain magnitude influences the number of replicates needed to satisfy statistical requirements, i.e. the number of colonies and fields used in the treatment and control groups.

Statistical considerations linked to the magnitude dimension

Comparing the arithmetic mean value of the colony strength in the control and in the treatment groups is not sufficient *per se*, as this difference may be due to chance (type I error). To tackle this, statistical tests with a pre-defined level of confidence are often used. The comparison of mean values is also the basis of the most common statistical tools used to evaluate these studies. The concept underlying these statistical tests is to check whether the difference between the means of the treatment and control groups is larger than the difference observed within each of the two groups. If so, then the difference is "flagged" as statistically significant, meaning that the observed difference between the groups is unlikely to be due to chance, with a probability reflected by the confidence level. However, lack of significance alone does not tell much about whether the magnitude dimension of

the SPG is met. The probability that a specific study will detect as significant a pre-defined difference between the mean values of the treatment and the control (i.e. the SPG magnitude) is defined as "power". If the power is low, then there is a high probability that a difference larger than the defined (SPG) magnitude will not be marked as significant (type II error). The power increases with larger magnitudes of the SPG and with higher replication (i.e. higher numbers of colonies and fields used in the treatment and control groups) in field studies. It follows that the selection of a certain magnitude will also drive the number of replicates needed in field studies in order to have a satisfactory power. This aspect is further discussed in section 3.4.

The **spatial scale** determines mainly the spatial distribution of the hives in the area used for the field study. In the EFSA bee guidance document[3], the identified spatial scale is the "edge of the field", which implies that all hives in a field study should be placed in the proximity of fields where the same crop is grown for both the treatment and the control group. In the treatment, the pesticide for which authorisation is sought is applied to the crop, while in the control group the crop remains untreated.

The **temporal scale** is the maximum time over which single or repeated exposure events are expected to exceed the acceptable effect level that can be tolerated. In principle, this includes the duration and the frequency of the effects, along with the interval between them. The temporal scale influences the frequency of the measurements and the length of the study. In particular, the EFSA bee guidance document[3] specified that field studies should last at least 2 brood cycles (about 42 days) as this was considered the minimum time to appropriately assess any potential adverse effect of pesticide. The EFSA bee guidance document[3] did not include any 'recovery option', but the entire SPG was based on a 'threshold option', thus a temporal scale for acceptable effects was not considered. For field studies, this means that the difference between the mean colony size in the treatment and the control should not exceed the magnitude threshold at any time. This presents practical limitations as the colony size cannot yet be measured continuously, as discussed in section 3.3.

The 'recovery option' and the 'threshold option'

These two options were first introduced for the pesticide risk assessment of aquatic organisms in EFSA PPR Panel (2013).

The 'recovery option' implies that transient effects above the threshold defined for the magnitude dimension may still be acceptable, if ecological recovery takes place within a defined time period.

The 'threshold option' implies that effects above the threshold defined for the magnitude dimension should not occur at any time.

3.2.2. Tiered approach and trigger values

Risk assessment does not uniquely rely on the reference tier (i.e. field studies) as this kind of experiment is complex and resource-intensive for all parties involved, including applicants and risk assessors. Risk assessment follows a tiered approach, starting from lower tiers that are typically based on simple more standardised laboratory studies and relatively simple exposure estimate approaches. Lower tiers are routinely used as a basis for screening substances in relation to particular concerns. In such lower tier laboratory studies, effects on bees are observed and recorded on an individual basis and not as colonies as in field studies.

Once the SPG dimensions are defined and it has been verified that these can be addressed in the reference tier, all different tiers of the risk assessment need to be calibrated accordingly.

Such a calibration exercise entails several steps, which allow linking standard endpoints such as L(D)D_{50}[7] to a reduction in colony size (SPG attribute of the identified ecological entity) equivalent to the acceptable effect (SPG magnitude) for a temporal scale defined on the basis of the exposure length in the

[7] lethal (daily) dose for 50% of the tested individual bees. Typical endpoint from laboratory studies with bees.

laboratory study (i.e. acute and chronic). The calibration, performed once all the SPG dimensions are defined, results in the definition of trigger values.

For the actual lower tier risk assessment, a risk quotient is calculated from the ratio between the dose equivalent to the standard laboratory endpoint (e.g. $L(D)D_{50}$) and the exposure predicted for the specific use of the substance, which accounts for the spatial scale of interest. The risk quotient is then compared to the trigger values described above. Hence, trigger values can be considered as thresholds that, if not exceeded by the risk quotient, guarantee the respect of the SPG. If trigger values in lower tiers are not exceeded, no further investigation is necessary, whereas if they are exceeded, higher tier risk assessments may be needed to further investigate whether the SPG is met.

3.3. SPG dimensions with Approach #2

As described in the first supporting supporting document[1], approach #2 is based on the analysis of the background variability of honey bee colony size. The analysis aims to define an **operating range (OR)**, i.e. the range of honey bee colony size given by their background variability[8]. The term "background" reflects that the analysis does not consider exposure to pesticides (e.g. like "controls" in experimental field studies).

Approach #2 does not require a complete revision of the SPG, i.e. a revision of all 5 dimensions (i.e. ecological entity, attribute, magnitude of the effects, spatial scale, temporal scale) implemented in the EFSA (2013). By selecting this approach, the MSs implicitly confirmed that the **ecological entity** is the **colony** and that the **attribute** is the **colony strength**.

The **spatial scale** implemented in EFSA (2013) is the **edge of field**. This means that the exposure estimation considered uniquely the colonies that are located at the edge of treated fields, i.e. those colonies that are likely to be most exposed among the ones in the area of use of a certain pesticide. The colonies living in the remaining hives (farther away from fields) are thus automatically protected. While in principle the exposure estimation could explicitly include all colonies (also the ones far from the treated fields), this has severe limitations in its practical implementation in the risk assessment. The level of exposure is likely to be influenced, among other things, by the distance between the hives and the treated field(s). Since the actual location of all bee colonies in Europe relative to agricultural crops is unknown (and likely not constant in time), implementing this approach in the risk assessment is unlikely to be feasible. The edge of field is the common spatial scale in the risk assessment for non-target organisms. This was also explained in Appendix A of the first supporting document[1].

By selecting approach #2, the risk managers implicitly agreed to revise mainly the definition of the current **magnitude** of effect and to implement a suitable **temporal scale** for the higher tier studies.

As reported in Table 1, in the EFSA bee guidance document[3], the **magnitude of effect** was agreed as 'negligible' and it was defined based on experts' judgement as colony size reduction < 7%, more specifically in the range of 3.5%-7%.

The EFSA Scientific Committee (2016) suggests avoiding using the terms 'negligible', 'small', 'medium', 'large' as descriptors of the magnitude of effects because these terms can be considered vague and qualitative.

The experts in the Working Group drafting the EFSA bee guidance document[3] unanimously agreed that '*a proportional reduction in colony size of greater than one-third would be likely to compromise the viability, pollinating capability and yield of any colony; this consideration was used to define an effect as 'large''*. This definition is generally accepted and not questioned, as it is rooted in a clear biological threshold (i.e. colony viability).

The current quantitative definition of "negligible effects" is also based on valuable experts' judgment. However, in contrast to the definition of "large" effects, assigning boundaries to this qualitative effect

[8] The object of the analysis was initially referred to as "natural variability". Following some relevant comments from MSs, EFSA changed the terminology to "background variability". This was done to clarify that the focus is not on colonies living in wild conditions. On the contrary, the focus is on managed honey bee colonies, like those that are likely to be used in field studies

class may be disputable, as it is not rooted in any clear biological threshold. Therefore, any attempt to quantitatively define "negligible" may lead to a controversial outcome, as demonstrated by the debate that occurred regarding the implementation of the EFSA bee guidance document[3].

The quantitative definition of the intermediate classes for "medium" and "small" effects were arbitrarily set at even intervals in the range between "large" and "negligible", but cannot be substantiated further.

The term "threshold of acceptable effects" was introduced with approach #2 because it is difficult to establish consensus on an undisputable scientific definition of qualitative class effects such as "negligible", "small", and "medium". Furthermore, the term "acceptable" is also in accordance with Annex II, point 3.8.3 of Regulation (EC) 1107/2009. Therefore, the concept of "acceptable effect" is considered as more suited in this context than any qualitative definition of effect class.

With approach #2, the magnitude of the effect on colony size is informed by the expected background variability (see section 3.3.1).

With approach #2, no explicit consideration is given to the temporal scale of the assessment, as the operating range is quantified in a continuous manner. In principle, this can be interpreted as an indication that the **temporal scale of acceptable effects** is not relevant, since any possible effect following the exposure to a pesticide should remain at a level indicated as acceptable at **any time.**

In practice, a temporal scale may be defined on the basis of practical limitations in the field studies (i.e. the reference tier). A continuous measurement of the colony size is not practically feasible yet, nor is it advisable to inspect the colonies too frequently, as this creates stress for the bees which would affect the results of the experiments. Until less invasive techniques become available, it is good practice to inspect the hive no more often than **every week** (see EPPO, 2010). Hence, in the time between two monitoring points, possible transient effects greater than the defined threshold could occur without being measured; however, the SPG can be considered met if the threshold of acceptable effects is not breached at the two monitoring time-points.

An alternative possibility, based on the biology of bees, could be to set the relevant temporal scale of acceptable effects as equal to **one honey bee worker brood cycle** (21 days). This is because it may be considered acceptable to have transient effects if these are compensated by the new generation of worker bees. However, this possibility should be carefully considered because if, for example, the transient effect over the 21 day occurs during the flowering period of the treated crop, pollination of the crop may be affected.

Even if temporal scale may, in practice, be part of the SPG definition, it will not have an impact on the calculation of the trigger values.

3.3.1. Informing the definition of the "magnitude" dimension using the concept of operating range (OR)

As explained in section 3.2, the SPG dimension related to the magnitude of acceptable effect can be directly measured in the reference tier (i.e. field studies) by comparing the mean colony sizes of the treatment and control groups. Effects are considered acceptable only if the mean colony size of the treatment group is not decreased by more than the magnitude dimension of the SPG, which is calculated relative to the mean colony size of the control group. Thus, the mean colony size in the control group should be taken as the reference.

The OR estimated with approach #2 considers uniquely colonies in the control group. The relative difference between the mean colony size and the lower limit of the OR informs on the maximum difference that can be caused by background variability. As such, the relative difference between the mean and the minimum, or any value between these, can be used to inform the definition of the magnitude of the acceptable effect of pesticides on colony size.

In summary, the following aspects should be considered:

- In the present analysis, honey bee colony dynamics are simulated over one year using the BEEHAVE model (see more about the use of this model under section 4.1).

- Simulations were carried out for 500 replicate control colonies in each of the considered scenarios (see more about the scenarios under section 4.2) allowing for assessing variability in size.
- The OR may consider the full variability range (hereafter "**full operating range**" or **FOR**), or it could be "restricted", by selecting a narrower range (hereafter **"restricted operating range**" or **ROR**), which excludes the colonies with lower size. The narrower the ROR, the smaller is the difference between the mean and the lower limit of the OR. Hence, the narrower the range, the smaller the magnitude of the acceptable effect.
- The part of the OR relevant for approach #2 is only the one below the mean, i.e. colonies that present a lower size compared to the mean. The part of the range above the mean is never considered in this approach, because there is no interest in imposing a limit on a beneficial effect, i.e. increase in colony size.
- The results are presented in terms of average variability over the entire simulated year, along with average variability over each season (spring: March-May; summer: June-August; autumn: September-November). The variability over winter was not considered in isolation, as measurements of colony size during this season are generally not performed.

Within this report, different operating ranges are defined by either the **percentage fractions of colonies retained** in the operating range or, which is equivalent, by the **percentiles of the variability** used as lower limit. For instance, when a fraction of 95% of colonies are retained in the restricted range, the lower limit would correspond to the 5th percentile of the full operating range.

The resulting difference between the mean of the colony size and the lower limit of the OR – irrespective of this being a FOR or a ROR – is always referred to as a **percentage difference**. The concept is graphically illustrated in Figure 1.

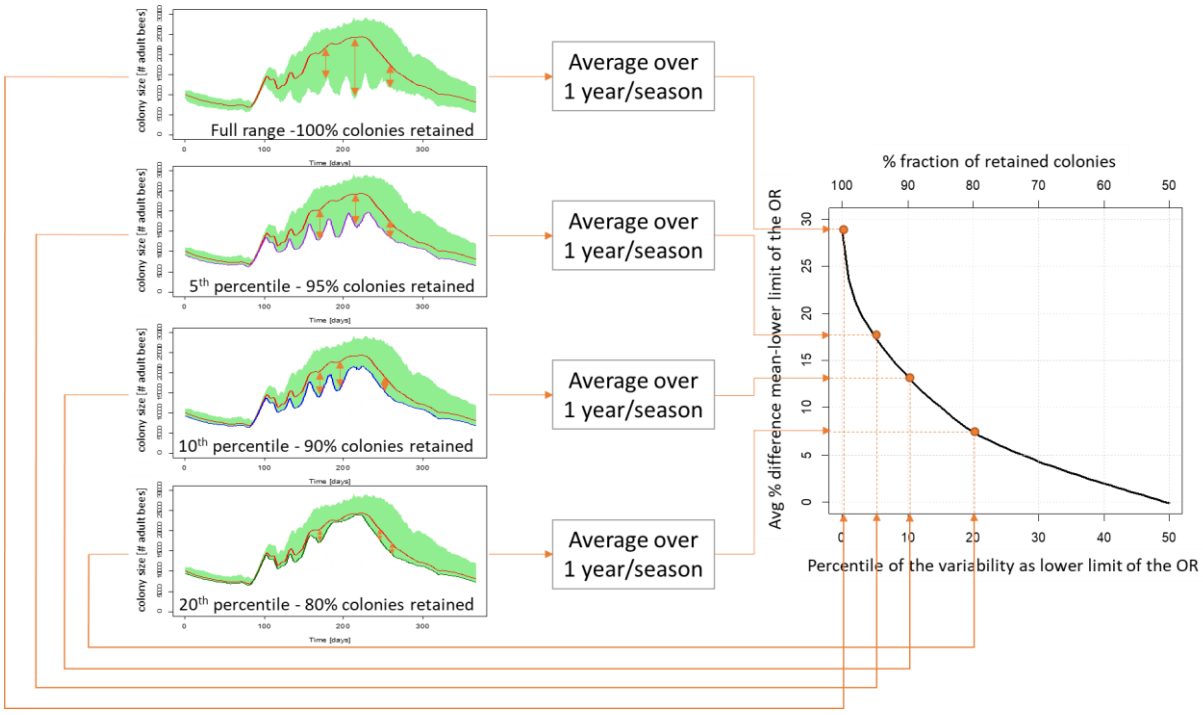

Figure 1: The full operating range (FOR) is depicted in the upper-most plot on the left as a green area. The green area represents the variability of the number of bees in the individual model hives.
The variability area can be described by a percentile scale, whereby the value 50 is assigned to the median and the value 0 represents the lower end of the variability distribution.
Setting the lower limit of the operating range to values higher than 0 leads to a corresponding exclusion of colonies from the operating range. Likewise, this leads to a decrease of the distance between the mean (red line) and the lower end of the operating range area (see green areas in the other plots on the left, i.e. restricted operating ranges). This distance is calculated

for every day of the year and then averaged over the entire year or over one season. The resulting average distance is expressed as a fraction of the mean value of bees (% of the mean).
The plot on the right side shows the effect of setting the lower limit of the operating range at values from 0 (no restriction) to 50 (all colonies below the median excluded):

- *the higher one moves up the percentile scale, the more colonies are excluded*
- *likewise, the average distance between mean and lower end of the operating range decrease from its maximum value at percentile 0 its minimum at percentiles close to 50.*

Summary box 2

Definition of the SPG

The level of protection is defined by 5 dimensions i.e. Ecological Entities, Attribute, Magnitude, Temporal and Spatial scale, with a high degree of certainty in the case of pesticides.

The dimensions were defined in the EFSA PPR opinion (2012) and implemented in the EFSA (2013) after risk manager consultation.

Principles of Approach #2

It does not require a full revision of the SPG implemented in the EFSA (2013), but could be limited to the dimension "magnitude of effects" and "temporal scale".

Magnitude dimension

The concept of "acceptable effect" was considered more suitable than any definition of effect class for the generic descriptor of "negligible", "small", "medium".

The analysis of the background variability in approach #2 informs on the definition of the magnitude dimension as the threshold of the acceptable effect on colony size reduction.

Temporal scale

Approach #2 does not consider explicitly any temporal scale, which can be interpreted as an indication that the threshold for acceptable effects (magnitude dimension) should not be exceeded at any time.

Practical limitation in the reference tier (i.e. field studies) suggests that more practical temporal scales can be based on either:

- The minimum interval between colony inspections (1 week)
- The length of a honey bee brood cycle (21 days)

Consequences for risk assessment

1) When effects of a pesticide observed in higher tier studies are above that threshold, the SPG is considered not met.

2) Since the lower tier risk assessment is calibrated to be compliant with the SPG, when the trigger values are breached, the SPG is not met.

4. Materials and methods

The analysis makes use of both modelling approaches and experimental data from literature and pesticide dossiers. For the modelling part, the BEEHAVE model (Becher et al. 2014) has been used. The experimental data from literature and pesticide dossiers are hereafter referred to as "external data", to highlight that these were produced independently of the model simulations.

Exploring the background variability of honey bee colony size using experimental data is in principle possible, but studies carried out with this scope are not readily available. In addition, experimental studies have other practical limitations:

1) colonies cannot be continuously monitored, and increasing the frequency of measurements also increases the level of stress to bees, altering the measured outcome;

2) the number of replicates that can be monitored is limited by the budget and other practical constraints of the study setup;

3) similarly, the possibility to analyse variability in different settings is subject to a big experimental effort, which requires significant investment in terms of time and economic resources.

In view of the limitations of the experimental approaches in isolation, EFSA considered that the task could be performed with the support of modelling. External data were used for calibration of the model and to check the plausibility of the final model predictions.

4.1. The use of the BEEHAVE model

The BEEHAVE model (Becher et al. 2014) simulates hive population dynamics by considering environmental factors, such as weather conditions, distance to flower patches and food availability. The model can also simulate the effects of infectious agents, like the *Varroa* mite and two associated viruses.

The model was evaluated by the EFSA PPR Panel in 2015[9]. The EFSA PPR Panel considered the conceptual model of BEEHAVE and the links between processes and variables logical and concluded that "*the validation of the BEEHAVE model for the original use fits quite well with the criteria required in the good modelling practice opinion (EFSA PPR Panel, 2014)".* The overall conclusion of the evaluation was that "*BEEHAVE performs well in modelling honeybee colony dynamics*".

On this basis, the EFSA WG has considered BEEHAVE the most appropriate model available for investigating the background variability of honey bee colonies in different environmental scenarios.

Nevertheless, the EFSA WG acknowledged and carefully considered the shortcomings identified by the PPR Panel[9]. The main limitation, i.e. that BEEHAVE is unsuitable for regulatory risk assessment, was deemed not relevant for the purpose of approach #2. This is because, in the analysis of the background variability of colonies, exposure to pesticides is not simulated, as risk from pesticides as a stressor is not evaluated.

Other limitations identified, which were considered relevant for the present exercise, have been fixed or mitigated. However, some other aspects could not be addressed within the scope and the timeframe of the current work. Possible sources of uncertainties related to those aspects are reflected in this document in section 6.

It is important to note that, following the evaluation of BEEHAVE in 2015, EFSA outsourced the development and validation of a mechanistic agent-based model (ApisRAM project), to assess risks to honey bee colonies from exposure to pesticides under different scenarios of combined stressors and factors (EFSA, 2016). Among the aims of ApisRAM there is an explicit willingness to overcome the limitations identified for BEEHAVE, particularly the lack of a pesticide module. In this view, the use of ApisRAM would provide benefits also for investigating the background colony variability as proposed in approach #2. However, ApisRAM is still under development, therefore it was not possible to propose it for the present exercise (see section 6.5 for more details on the use of ApisRAM).

Overall, in May 2020 the EFSA WG concluded that the use of BEEHAVE represented the best option currently available for the scope proposed with approach #2.

Summary box 3
Why analyse the background variability with the support of modelling?
• The practical limitations of field studies prevent a comprehensive analysis of the colony background variability, while this can be performed with the support of models simulating the colony dynamics (e.g. in-hive processes, feeding behaviours etc).

[9] EFSA PPR Panel, 2015

- The BEEHAVE model was evaluated in 2015 by the EFSA PPR Panel, who considered it suitable for simulating colony dynamics and therefore this model was selected for this analysis as the best available option.

4.2. Environmental scenarios

Honey bee colonies behave in different ways depending on the environmental context they are part of. As a consequence, the background variability in colony sizes can also vary, resulting in different operating ranges for different environmental contexts. In order to cover a realistic range of the different European conditions, EFSA superimposed a 5x5 grid over the map of the EU, leading to 25 cells of equal size. 5 grid cells only contained sea. For the remaining 20 cells, EFSA randomly selected one location per cell and attempted the construction of related environmental scenarios for running model simulations in each one of them.

The exercise was carried out for 19 of the 20 locations: for the northernmost location, close to Kittilä (Finland), no successful scenario calibration was achieved, probably due to the rather extreme climatic conditions north of the Arctic Circle.

To ease the interpretation, in this report results are presented for six scenarios[10]. These scenarios represent the highest and lowest colony size variability – calculated as mean coefficient of variation over the entire year – in each of the three regulatory zones. The variability in the other scenarios is generally intermediate between those reported here. The locations corresponding to the scenarios used in this report are illustrated in red in Figure 2.

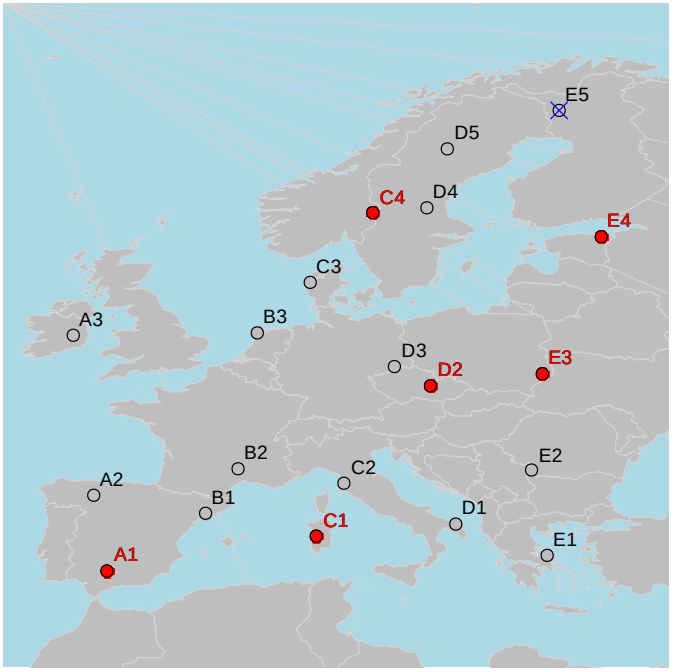

Figure 2: locations corresponding to scenarios used for the analysis of colony size variability. For scenario E5, no successful scenario calibration was achieved. The six locations marked in red are those corresponding to the scenarios reported in this document. They are the scenarios with the highest and lowest colony size variability in each of the three regulatory zones.

4.3. Description of the scenarios

BEEHAVE uses a large number of parameters to describe the complex interactions within the colony and between the colony and the surrounding environment. Most parameters were left unchanged with respect to the default setting used in Becher et al. (2014).

[10] Results for all 19 scenarios will be included in a technical report (in preparation). However, they can be provided at any time, in case Risk Manager would like to have a more in-depth assessment of the results.

However, several aspects were modified *ad-hoc* to define the scenarios. Some aspects were adjusted for each scenario in order to describe their specificity, while others were kept constant across the scenarios. A brief overview of the elements that were considered to define the environmental scenarios is reported in Table 2.

Table 2: overview of the elements describing the environmental scenarios used in the BEEHAVE simulations

Main area	Item	Scenario-specific	Description
Foraging / climate	Foraging hours per day	Yes	Temperature and irradiance (i.e. measure of sunlight) for the years 2005-2016 have been used to estimate the maximum number of foraging hours for any day of the year and for each scenario. For further details see section 4.4.1.
Landscape structure	Number of patches	No	A simplified landscape with two food patches has been used in all scenarios. This is the same landscape used in the original implementation of BEEHAVE.*
	Distance of the patches to the hive	Yes	Parameter calibrated for each scenario (see section 4.4.1).
Resource availability	Max availability of pollen and nectar	Yes	Parameter calibrated for each scenario (see section 4.4.1).
	Availability of pollen and nectar in time	Yes	Adjusted to the foraging period. This was done in order to ensure longer flower availability in climates that allow longer foraging windows.
Bee biology	Maximal egg-laying rate of the queen over time	Yes	Adjusted to the foraging period. This was done in order to indirectly link the egg-laying with climatic variables (i.e. temperature and sunlight) that are likely having a strong influence on the onset and the offset of this activity.
	Forager mortality rate	No	The mortality rate was calibrated by aligning the resulting forager lifespans (from age of first forage to death) to the empirical values retrieved from the literature. It must be noted that while the mortality rate was kept constant across the scenarios, the resulting lifespans were different. See section 4.4.1 for details.
Beekeeping practices	Amount of added fondant	Yes	Parameter calibrated for each scenario (see section 4.4.1).
	Honey harvesting period	Yes	Adjusted to the foraging window.
	Initial colony size	No	The average starting bee population in the simulated colonies was 10000 honey bees (±1000), which is in line with typical colony sizes used in field studies This is also in line with the findings from Harbo (1986), who found that a similar starting population size (9000 bees) was optimal for balancing brood and honey production efficiency.

*Due to limited data availability, a more realistic definition of landscape scenarios based on data was not possible. However, since the EFSA WG considered that the adopted simplification of the landscape

was a crucial point, a separate analysis has been set up to explore the effect of landscape complexity on the final outcome i.e. variability in colony size as simulated by the model. The results of this analysis are presented in Appendix B.

4.4. External data

External data, mainly retrieved from literature and pesticide dossiers, were used in two different phases of the work. In a first phase, literature data were used to calibrate the different scenarios. In a later phase, other data from pesticide dossiers were used to check the plausibility of the model simulations.

4.4.1. Data used for scenario calibration

Some of the parameters listed in section 4.3 have been adjusted for each scenario to reflect their specificity. The process of adjusting these parameters is hereafter called "calibration".

The maximal foraging hours per day have been calculated following an empirical relationship based on an experimental study from Vicens and Bosch (2000), which considers temperature and solar irradiance. Location-specific data for these two climatic variables were retrieved from a JRC GIS platform[11]. Different foraging hours per day have been calculated for each scenario. As expected, the foraging season was longer in the southern scenarios and shorter in the northern scenarios.

The availability of flowers during the year, and the presence of a latitudinal gradient in this respect, have been verified by checking several beekeepers' calendars (e.g. Leida et al.), which are available for some areas of Europe. Data from the scientific literature were also considered (e.g. Baude et al. 2016). However, it was not possible to find any empirical relationship to link the length of the flowering season to geographical features, so these data were only considered in a qualitative way.

In order to overcome this problem, the temporal availability of food (i.e. length of the flowering season) was adjusted, in a scenario-specific way, to the length of the foraging season. Plant phenology is known to be influenced by temperature and sunlight. By linking the flowering season to the foraging one, the former has also been indirectly linked to relevant climatic variables.

A similar approach was also adopted for the maximum egg-laying rate over time, which is known to be influenced by temperature and sunlight hours. The default egg-laying rate used by Becher et al. (2014) was based on a previous model (Schmickl and Crailsheim, 2007). Their model for egg-laying was obtained by fitting data from Bodenheimer (1937), which were based on observations from Ebert (1922). Considering the time of publication, which was 17 years before the insecticide properties of DDT[12] were discovered, it is reasonable to assume that bees were not exposed to synthetic pesticides in this study.

The forager mortality rate is implemented in BEEHAVE as a probability of dying per second spent outside the hive. The calibration in this case was not made scenario by scenario, but considering all scenarios at once. This is because the same probability of dying results in a different forager lifespan in each scenario. The reference data used for the calibration are those included in the recent review of the evidence of bee background mortality[13]. Within this review, data from both agricultural and non-agricultural areas were considered. Studies presenting evidence that bees were exposed to insecticides were nonetheless always excluded, as laid out in the protocol drafted before the review was performed. The calibration performed in the present work aimed to align the modelled forager lifespan, i.e. the time from the age of first forage to death, to the empirical values retrieved from the literature. The outcome of the calibration resulted in a considerably lower probability of dying than the one used in the default model implementation: from 1E-05 used in Becher et al. (2014) to 3.84E-06 used in the present work.

Maximum food (i.e. pollen and nectar) levels in the flowering patches, distances from the patches, and the amount of added fondant all combine to determine the energy balance of the colonies. As such, these were calibrated together by considering multiple sources of data. Since these parameters were not available for many different parts of Europe, the values in the different scenarios were set to match

[11] https://ec.europa.eu/jrc/en/pvgis
[12] dichloro-diphenyl-trichloroethane
[13] EFSA et al. (2020a)

country-specific average honey yields. To this purpose, data were collected from FAOSTAT (year 2010-2018), from the EU National Apicultural Programmes of the EU Commission (2020; referred to year 2017-2018), from Chauzat et al. (2013, referred to year 2010). In addition, data from the Prevention of Honey Bee COlony LOSSes (COLOSS) project (Hatjina et al. 2014) were also consulted.

Typical pollen:nectar ratios for several plant species reported in the literature have been consolidated in tables presented in Becher et al. (2016) and Agatz et al. (2019).

Food levels were also calibrated by accounting for plausible colony size in the different seasons, for example the maximal colony size in summer. A qualitative consideration of data on colony size in the different areas was obtained again from the COLOSS project (Hatjina et al. 2014), and partially from a beekeeper survey carried out by EFSA in 2020[13].

4.4.2. Data used for checking the plausibility of the model simulations

Data on colony size have been extracted from control colonies of 32 field studies. In particular, the data set assembled for the review of the neonicotinoids[14] through an open call for data and systematic literature search, in addition to other pesticide dossier studies, were used for the present analysis. About 90 studies had been initially considered, but many were excluded due to either lack of details or reliability issues (e.g. application of insecticides also in the area of the control colonies, evidence of control contamination, etc.). Overall, about 2000 colony size values have been extracted from 297 time points. The variability among replicates in these control colonies was used as a reference to check the plausibility of the model simulations.

Summary box 4

Model calibration/Environmental scenarios

- To cover a realistic range of the different European conditions, EFSA selected several locations in the EU with a semi-randomised procedure.
- Environmental scenarios were set up for each of the selected locations. These environmental scenarios were used for running model simulations.
- The process of setting up the scenarios required definition of parameters, some of them scenario-specific, others kept constant across scenarios.
- The definition of these parameters was performed via a calibration of the model for each scenario.
- The calibration made use of literature data of different sorts.
- Other input parameters feeding the model were left unchanged with respect to the default setting used by the model's authors.
- Data from pesticide dossiers have been used to check the plausibility of the model simulations.

5. Results

The focus of the following sections is on the quantification of the colony size variability, expressed as relative percentage difference between the mean and the lower limit of the OR. Together with the full operating range (FOR) several restricted operating ranges (RORs) are presented, and these are indicated by: 1) the percentage fraction of colonies retained in the OR and by 2) the percentile of the variability used as lower limit of the OR.

As mentioned earlier, simulations were performed for 500 replicates per scenario. The results are presented in terms of average variability over the entire simulated year, along with average variability over each season (spring: March-May; summer: June-August; autumn: September-November). The variability over winter was not considered in isolation, as measurements of colony size during this season are generally not performed.

[14] EFSA, 2018

5.1. Overall results of the simulations

A summary of the entire simulation exercise for the six scenarios together is presented in Table 3. The values represent the minimum and maximum percent differences between the mean colony size and the lower end of the OR based on the lowest and highest simulated background variability across all the scenarios. These values can inform on the definition of the acceptable level of the colony size reductions to be specified in the SPG.

Table 3: percentage difference between the mean colony size and the lower limit of the OR. Values are presented as the minimum and maximum across the six scenarios. The OR is presented as the whole variability (i.e. the FOR) and as "restricted" variability ranges (RORs) to various extents.

Percentile of the variability as lower limit of the OR	% fraction of colonies retained in the OR	% difference between the mean colony size and the lower limit of the OR			
		Full year (min-max)	Spring (min-max)	Summer (min-max)	Autumn (min-max)
Whole range (FOR)	100%	20.3% - 31.1%	18.7% - 25.4%	12.8% - 47.1%	20.7% - 44.5%
5th percentile	95%	9.9% - 17.9%	9.5% - 14.8%	6.1% - 26.4%	11.5% - 27.2%
10th percentile	90%	7.3% - 13.3%	7.1% - 12.0%	4.8% - 18.3%	8.2% - 19.4%
20th percentile	80%	4.8% - 8.6%	4.5% - 8.4%	3.2% - 9.6%	5.2% - 11.1%
30th percentile	70%	3.0% - 5.7%	2.6% - 5.4%	1.5% - 6.1%	3.3% - 8.2%
40th percentile	60%	0.8% - 2.9%	1.1% - 2.6%	-1.0%* - 2.8%	0.3% - 5.7%
50th percentile	50%	-1.4%* - 0.2%	-0.3%* - 0.0%	-3.1%* - 0.0%	-2.8%* - 3.1%

* Value > mean, should not be considered for threshold derivation

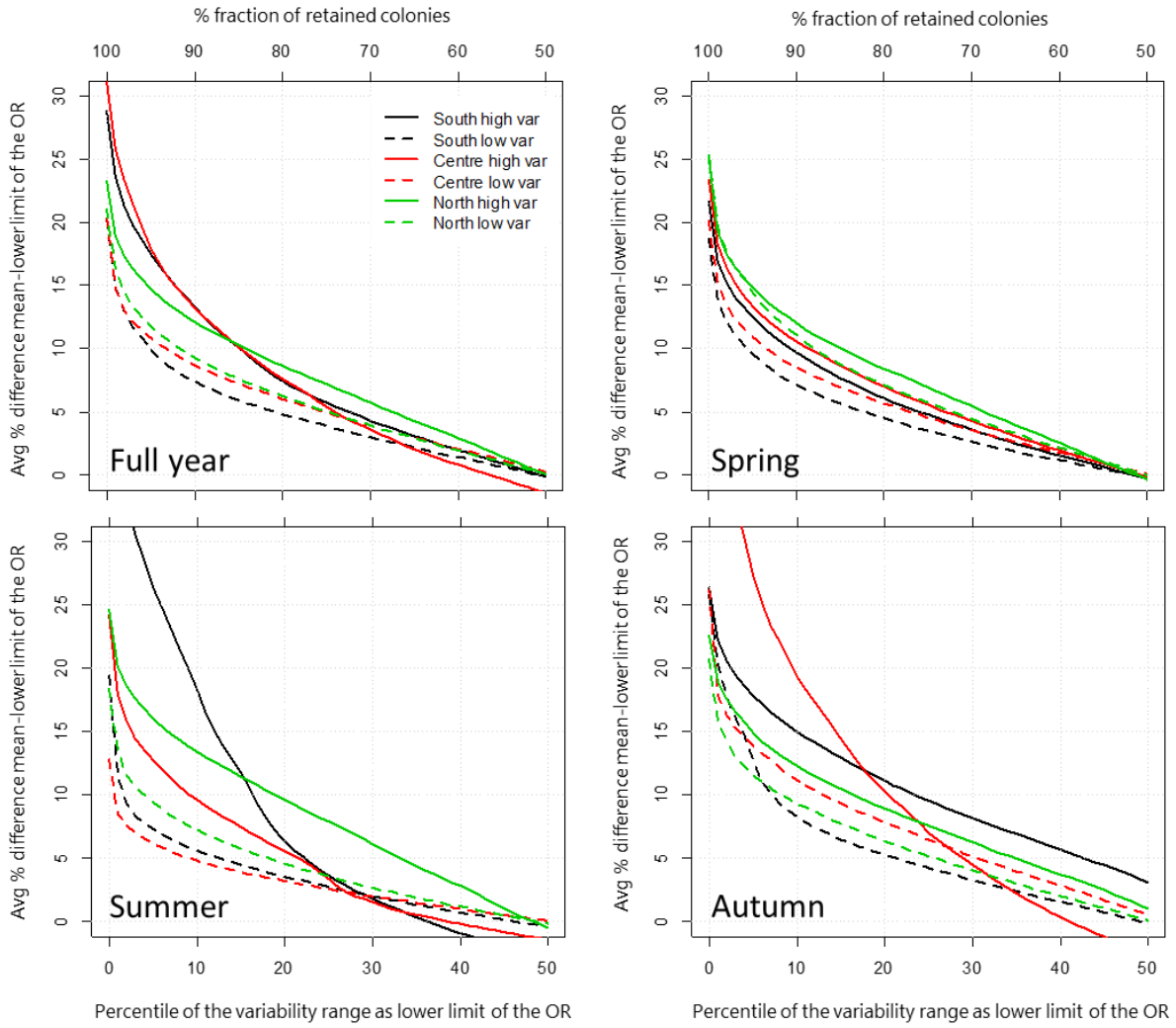

Figure 3: Comparison of the variabilities between scenarios, over the full year, spring, summer, and autumn

While the simulations revealed no striking differences between the three regulatory zones, generally the variability tended to be slightly higher in the northern scenarios during spring (see Figure 3, top-right). In summer the variability was very high in one of the two southern scenarios, but not in the other (see Figure 3, bottom-left). A similar pattern was observed for the central zone in autumn (Figure 3, bottom right).

The variability peaked in spring/summer in the northern scenarios, in autumn in the central scenarios, while it was more heterogenous in the southern scenarios (one peaked in summer, the other in autumn).

5.2. Interpretation of the results

The results presented in section 5.1 cover the simulated variabilities of 500 "control" colony replicates for the whole year and in spring, summer and autumn. In Appendix A the results are presented also per regulatory zone.

The whole range of the simulated variabilities (FOR) includes the "stronger" control colonies i.e. the colonies at the top of the range and the "weaker" ones, i.e. colonies at the bottom of the range. The mean of the range represents the mean variability of the 500 "control" colony replicates.

The colonies showing sizes lower than the mean are the relevant ones for approach #2. This is because the attribute to protect is the colony strength (=colony size). Detrimental effects in the reference tier (i.e. field studies) are expected to cause the mean size of the treated colonies to be reduced compared

to the mean size of the controls. Hence, the part of the range which is below the mean is relevant for defining the magnitude of the effect. Identifying a threshold for the magnitude of acceptable effects in the range above the mean would mean that the effect is acceptable only if it causes the mean colony size in the treatment to be higher than the mean colony size in the control, which is nonsensical in a risk assessment context.

The explicit consideration of the honey bee colony size background variability clarifies the extent to which colonies unexposed to pesticides can deviate from each other, and, most importantly, how much they can deviate from their average size.

The results in Table 3 and in Figure 3 show the percentage of colony size "reduction" which correspond to a fraction of colonies retained in the OR. These percentages of size "reduction" are calculated as difference between the mean colony size and the lower limit of the OR. The results can also be read as selecting a ROR by defining a given percentile of the colony size variability distribution as the lower limit of the ROR.

Retaining in the ROR 95% of the total colony size variability (lower limit at the 5th percentile of the FOR) or the 90% (lower limit at the 10th percentile) is equivalent to excluding, for the definition of the magnitude of acceptable effect, the 5% or the 10% weakest colonies. If these restrictions are selected to determine the threshold of acceptable effects when this threshold is used to evaluate a pesticide, it means that the mean colony size of the exposed colonies in the treatment group should always be larger than all the 5% or 10% most vulnerable colonies in the control. In other words, the treatment colonies must, on average, perform better than the 5-10 % weakest control colonies, in order to meet the SPG.

It has to be noted that the exclusion of only 5% of the colonies would significantly reduce the range of the background variability, and thus the threshold for acceptable effect on colony size reduction. By taking as example the scenario with the lowest variability over the entire year, the threshold would be 9.9%, instead of 20.3% without restriction. In the same scenario, when 10% of the colonies are excluded, the range of the variability is only slightly reduced, and the threshold would be 7.3% and so on. It is clear that, the narrower the range, the smaller – and thus more conservative – is the magnitude of the acceptable effect.

Summary box 5
Percentages of colony size reduction
Threshold(s) of acceptable effects represent percentage(s) of mean colony size reduction.
A threshold of X% (for example 20.3%, or 9.9%, or 7.3%, or 4.8% etc) indicates that the mean size of colonies exposed to a pesticide (i.e. the treatment) should not be lower by more than X% (for example 20.3%, or 9.9% or 7.3%, 4.8% etc) compared to the mean size of the colony in the control.

5.2.1. Recommendations on how to interpret the results

The background variability and the analysis of its distribution should support the selection of a threshold of acceptable effects. Further considerations are necessary on how to use this analysis for determining the threshold of acceptable effect. In particular, the following is recommended:

- **Colony size reductions of one third were already identified (see section 3.3) as a threshold potentially leading to the impairment of the colony viability.** The results of the simulations show that, in some circumstances (see Table 3), the distance between the mean size and the lower limit of the OR (either FOR or ROR) is close to or even higher than one third (i.e.33%).
- **By using a more restricted OR the weaker colonies are left out, and therefore the threshold of acceptable effects is more conservative**. The results of the simulations show that only a few colonies are substantially weaker than the average colony size, even when no exposure to pesticide is considered. By restricting the OR for threshold derivation, these weaker

colonies will not be used as a reference for acceptable effects, resulting in a general higher protection level.

- **A restriction to the 50th percentile of the variability should not be considered for the threshold derivation**, since this would mean that, in many cases, only beneficial effects of pesticides are considered acceptable i.e. increase in colony size compared to the control.
- **The threshold of acceptable effect should be implementable in the reference tier**; since, currently, the reference tier is represented by the field studies, it should be realistically measurable in those studies (see section 7).

More considerations about variability distributions in risk assessment and selection of relevant thresholds can be found in Appendix C.

Additional considerations for selecting the OR restrictions/or the threshold of acceptable effect can be found in section 8.

5.3. Plausibility of the model simulations

As previously mentioned (section 4.4.2), colony size data have been extracted from control colonies of 32 field studies. These data were not used in the calibration phase, so they are fully independent from the model simulations. Almost all field studies were conducted in the central zone. Five studies were formally conducted in the Southern zone, but the location of three of them (Alsace and Champagne) is probably more representative of central European conditions. None of the studies was carried out in the Northern zone. Starting from those, the variability in size between replicate colonies at every measurement time point has been quantified as coefficient of variation (CV), i.e. the standard deviation divided by the mean.

Overall, these experimental data show two very clear trends:

1) The variability of the CV among studies is extremely high. This can be considered as the "variability of the variabilities" and shows that, while in some studies the colony size was relatively similar among replicates, in some other studies this was not the case.
2) Overall, the CV tends to increase with time (see Figure 7). This is likely to happen because, at the beginning of field studies, colony sizes are often purposefully equalised, in order to have more meaningful comparisons.

The present modelling analysis simulates colony dynamics for one year. Hence, only data related to the first year of each study were retained. Furthermore, studies presenting a CV > 0.3 at the start of the test period were excluded, as this rather high variability was already there before the colonies were placed in a common environment. Having too high variability (i.e. CV > 0.3) would have resulted in a non-plausible comparison with the model results which started with a CV ≈ 0.1 at the beginning of the simulation. This led to the exclusion of 87 data points, while 210 were kept for the analysis.

The comparison of the variability from field studies and from the simulated scenarios is presented in Figure 7. The variability across all the available control replicates from the 32 field studies is represented as CV at certain time together with a median tendency in time. The simulated variabilities are shown as average, minimum, and maximum CV over the three seasons and the different regulatory zones.

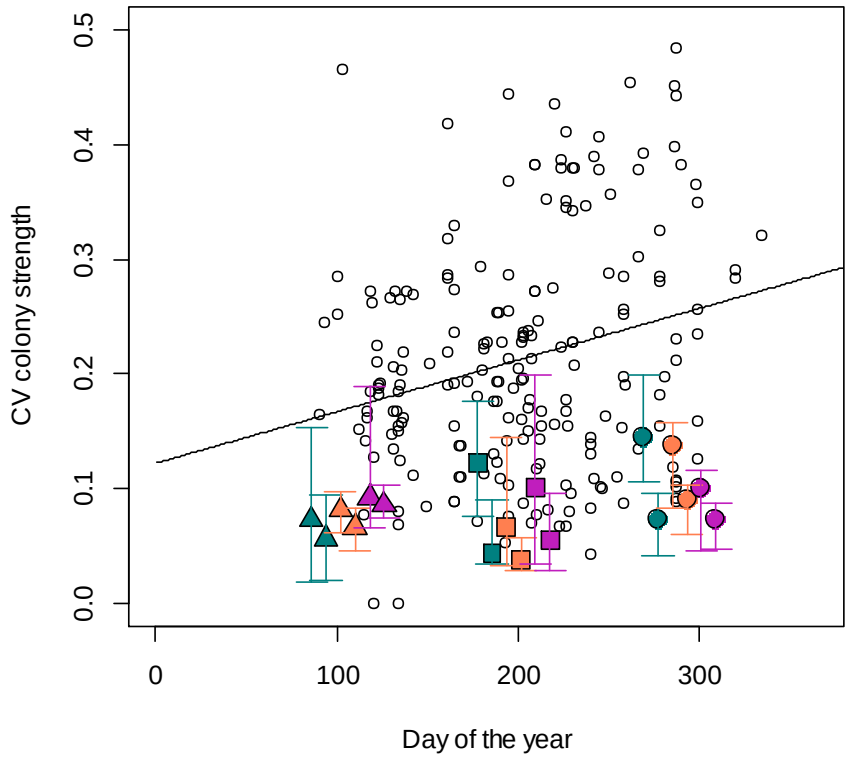

Figure 4: comparison of the variability from field studies (open circles) and from the simulated scenarios with BEEHAVE (coloured symbols). The variability is quantified as coefficient of variation (CV) across all the available replicates. While the experimental points express measured variability at a certain time, the values related to simulations express the average (solid symbols), minimum (lower whiskers), and maximum (upper whiskers) CV over the three seasons (triangle=spring, square=summer, circle=autumn). Different colours indicate different regulatory zones (green=south, orange=centre, purple=north) and both the low and the high variability scenarios for each zone are included in the figure. The black solid line expressed the median tendency for the experimental variability in time.

In general, the variability simulated by the model was smaller than the median variability estimated on the basis of the field studies. While the simulated variabilities are generally in the range of the experimental ones, they tend to be in the lower part. Since the simulated colony variability will be the basis for defining acceptable effects, risk managers should consider that an underestimation of the variability is more conservative than an overestimation.

Summary box 6
Plausibility of the model predictions
The variability simulated by the model was smaller than the median variability observed in field studies, but still in the range of plausible values. An underestimation of the variability is more conservative than an overestimation.

6. Uncertainties and potential future developments

As mentioned above, despite relying on the best available science, the simulations presented in this document have several uncertainties. These are related to the intrinsic limitations of predictive models (e.g. simplification of complex natural processes), to specific limitations identified for BEEHAVE, to the relevance of some of the input values for the present analysis, and to the definition of the environmental scenarios.

The influence of some uncertainties was quantified to provide the most comprehensive information to risk managers for the final SPG definition. In particular, the influence of landscape complexity was quantified (see Appendix B).

For other uncertainties, however, such quantification could not be performed within the scope and the timeframe of this exercise. Performing some of these analyses would require profound modification of the model, which is considered outside the scope of the current exercise. Therefore, for those "unquantified" uncertainties the potential influence on the outcome is unknown.

6.1. Limitations of BEEHAVE identified in EFSA PPR Panel (2015)

Among the limitations identified in EFSA PPR Panel (2015), the main one relates to the use of BEEHAVE in the regulatory risk assessment of pesticides. Considering the use of BEEHAVE (see also section 4.1) in the present document, this limitation is deemed neither relevant nor applicable for this analysis. For example, the lack of a pesticide module is unimportant when the simulations are run to address colony dynamics where pesticide exposure is absent.

However, other limitations identified by the EFSA PPR Panel (2015) are still relevant and have the potential to influence the outcome of the analysis presented in section 5. Specifically, some of the default parameters (e.g. food consumption, flight velocity, maximum egg laying, etc.) used in the model are not fully supported by empirical data.

6.2. Limitations of BEEHAVE identified in the present analysis

Additional limitations intrinsic to BEEHAVE for the specific purpose of the simulations have been identified while performing this analysis.

The first aspect regards how foraging intensity is simulated. In the real world, food needs and environmental conditions are both important drivers for the number of bees leaving the hive to forage (i.e. foraging intensity). However, in BEEHAVE, environmental conditions (i.e. temperature and sunlight hours/irradiance) do not determine foraging intensity but are used only to quantify the daily foraging time-window. The influence of this on colony size variability is, at the moment, not quantifiable without introducing significant modifications to BEEHAVE.

Perhaps the most relevant aspect is that the variability between replicate runs in BEEHAVE is determined by stochastic (=random) processes described by probability distributions. Stochastic is defined as an event determined by random processes. This means that some of the parameters of the model are not fixed, but assume different values at every run, under equal conditions, on the basis of probability distributions. The variation in the value(s) of these parameters determines the variability in the model output.

In BEEHAVE, stochasticity concerns two main processes – mortality and forager activity:

- Mortality of single bees is random in BEEHAVE. However, this occurs with different pre-defined probabilities. In-hive bees die with different daily probabilities for each life stage (eggs, larvae, pupae, adult, all different for drones and workers). Scarcity of nurse bees or lack of pollen may influence brood mortality as well, but these aspects are not random and they are added on top of the stochastic mortality. Foragers have a certain probability of dying for every second spent foraging, so that longer foraging times entail higher mortality probability.
- Forager activities have several random aspects: the choice of a bee to start or stop its foraging activity, the choice to forage pollen and/or nectar are all determined by probabilities, which are in turn driven by the hive needs. The detection of a flower patch occurs with a certain probability, driven in our analysis by the patch size and its distance from the hive.
- Other stochastic parameters exist for the *Varroa* module of BEEHAVE, which was not used in the present simulations (see section 6.4).

While the stochastic approach *per se* can be considered reasonable, there seems to be a lack of explicit justification for the probability distributions chosen in the model. Also in this case, the influence of this aspect on the colony size variability was not quantified.

6.3. Relevance of input values for the present analysis

The objective of the present work is to illustrate an analysis of colony strength background variability for a perfect control in the reference tier of the risk assessment (i.e. a field study). This entails two main aspects:

- the conditions surrounding the simulated colonies should resemble the typical habitat for honey bees in agricultural areas, where field studies are carried out.
- a complete lack of exposure to pesticides, since this should serve as a benchmark for effects of pesticides.

In the real world, agricultural areas are unlikely to be completely pesticide-free, while this could be the case for some non-agricultural areas (e.g. mountains, forests). However, habitats in these areas are completely different in terms of structure, food availability, competition and predation compared to agricultural areas.

Some aspects related to the habitat structures (i.e. food availability etc.) can be dealt with directly in the model by adjusting some of the scenario parameters. Limitations in the simulated habitat structure, also in terms of food availability, are discussed in detail in the next section.

BEEHAVE makes use of input values, i.e. not calculated by the model but imposed by the user, describing the biology of bees. These encompass aspects related to reproduction and development (e.g. max egg laying, length of each life stage, etc.), foraging (e.g. flight velocity, maximum amount of pollen and nectar carried by one bee, time needed for food collection and unloading, etc.), food consumption, mortality (probability of death for in-hive for adult and brood, foraging mortality etc.), and brood care (e.g. maximum amount of brood nurse bees can care for).

Some of these input values are known to be robustly estimated: for example, the developmental time of eggs, larvae, and pupae are known to be the same in agricultural areas and in the laboratory. However, at least in principle, each one of the aforementioned biology-related input values can be influenced by both the habitat type and exposure to pesticides (or to any other hazardous chemical). Hence, the conditions of the experimental studies used to derive these input values are relevant.

For mortality of adult bees, the input values were calibrated on the basis of the data included in the recent review of the evidence of bee background mortality (EFSA et al, 2020a). For that review, data from both agricultural and non-agricultural areas were considered, but studies presenting evidence that bees were exposed to insecticides were excluded.

The maximum egg-laying rate over time was adjusted for each scenario considering daily temperature and sunlight hours. However, the starting point was the egg-laying rate used by Becher et al. (2014), which was based on a previous model (Schmickl and Crailsheim, 2007). This model made use of observations from Ebert (1922). While it was not possible to ascertain whether these observations were performed in agricultural areas, considering the time of publication it is safe to assume that bees were not exposed to synthetic pesticides in the original study.

The origin of all other input values had not been investigated in depth, but many were derived from rather old studies, whose level of detail does not allow to ascertain whether they were carried out in agricultural areas or not.

6.4. Uncertainties in the scenario definition

The main limitations of the definition of the scenarios is related to the lack of suitable ready-to-use data. While some aspects were satisfactorily addressed (i.e. climatic conditions, average mortality rate of foragers) by using available data, others relied on indirect estimations, which entailed relevant uncertainties in the final outcome.

The main source of uncertainty in this sense is related to food (i.e. pollen and nectar) availability. While some data are available in the literature (Becher et al. 2016, Agatz et al. 2019, Baude et al. 2016, Timberlake et al. 2019) these are too scattered to establish reliable food levels in all the scenarios.

Hence, nectar levels in the landscape were indirectly calibrated by aligning as much as possible the simulated amount of harvested honey with available data on average honey yield (FAOSTAT; European Commission, 2020; Chauzat et al., 2013). However, these available data are averages at country level and might not be representative for the specific selected locations. Hatjina et al. (2014) reported site-specific data for two years and several European locations, but: 1) these locations did not overlap with the scenario locations used in this report, and 2) extrapolation was hampered by the lack of a clear geographical pattern.

Realistic pollen levels were also indirectly estimated by using experimental nectar:pollen ratios available in the literature (Becher et al. 2016, Agatz et al. 2019).

The scenario calibration exercise also considered data on maximum colony size. Hatjina et al. (2014) reported information for a larger area, but no consistent geographical pattern could be identified in the two years of the study. Finally, the results from the beekeeper survey proved to be of limited reliability (as already noted in EFSA et al., 2020a) with a very limited coverage of southern countries and several unrealistically high values indicated (e.g. >1 million bees per hive). Overall, the available information on average colony size could only be used qualitatively to check that plausible values were simulated, but not to calibrate differences between scenarios.

The influence of adopting a simplified landscape description was analysed systematically (see Appendix B). The results of this analysis show that landscape complexity has the potential to increase colony strength variability, mainly via increasing food inflow variability. Consequently, the use of a simplified landscape context, as in the model standard settings, may have resulted in an underestimation of the variability.

BEEHAVE offers the possibility to simulate the effects of *Varroa* mite and two associated viruses: deformed wing virus (DWV) and acute paralysis virus (APV). However, in the present analysis, it was decided not to include a simulation of pathogen infections related to *Varroa*. This decision was taken mainly for two reasons:

- Several concerns were raised by the EFSA PPR Panel in their 2015 statement about this module. The PPR panel concluded that the impact of *Varroa*/viruses on colony survival was underestimated, and that the simulated treatment against *Varroa* might be too effective to be realistic.
- The activation of the *Varroa* module in BEEHAVE entails the selection of many additional parameters (e.g. number of initial mites, initial infection rate, choice between different mite reproductive models, choice between the simulated viruses, etc.), which could be different for each scenario. In the absence of relevant data on these parameters, it was considered that the uncertainty introduced by simulating *Varroa*/viruses may potentially exceed the uncertainty of simulations that explicitly ignore this aspect.

Exploratory simulations were nevertheless performed. They revealed that the activation of the *Varroa* module is likely to increase the colony size variability between replicates. Hence, risk managers should be aware that the choice of not including *Varroa*/virus in the simulation may have resulted in an underestimation of the variability. An underestimation of the variability is more conservative than an overestimation.

Summary box 7

Uncertainties/limitation of BEEHAVE

- Any model entails a simplification of complex natural processes.
- The lack of a pesticide module is unimportant when the simulations are run to address colony dynamics where pesticide exposure is absent.
- Some of the default parameters (e.g. food consumption, flight velocity, maximum egg laying, etc.) used in the model are not fully supported by empirical data.
- The source of variability between replicate runs in BEEHAVE is determined by stochastic (=random) processes described by probability distributions.
- Landscape complexity has the potential to increase colony strength variability, mainly via increasing food inflow variability. Consequently, the use of a simplified landscape context, as in the model standard settings, may have resulted in an underestimation of the variability.
- The choice of not including *Varroa*/virus in the simulation may have resulted in an underestimation of the variability.

6.5. Outlook

As mentioned in section 6.4, many of the uncertainties are related to lack of reliable data with sufficient coverage of different European conditions.

Several ongoing research projects have the potential to fill some of the identified data gaps. For example, B-GOOD (until May 2023) and PoshBee (until May 2023) will collect information on multiple aspects connected to the health of bees (honey bees under B-GOOD and bumble bees and solitary bees under PoshBee). While the focus is mainly on the effects on bee health caused by multiple stressors, it is likely that relevant data about the landscape and other elements related to bee nutrition will also be collected.

Finally, as previously mentioned, EFSA outsourced the development and validation of a mechanistic agent-based model (ApisRAM project). In the context of the development of ApisRAM, data are being collected as input for modelling environmental scenarios (i.e. Denmark, Portugal) at a very high spatiotemporal resolution. This should include the farming activities and phenology of the vegetation.

With the support of the B-GOOD project, additional data will be made available for ApisRAM from 8 other countries including Belgium and the Netherlands (in 2023) which can be used to evaluate the model's performance. With the support of the university of Jagiellonion (Poland) and the PoshBee project, the university of Aarhus and Trinity College (Ireland) will develop bee population dynamics models for solitary bees and bumble bees (in August 2021 and May 2023, respectively).

The EFSA procurement for the development of ApisRAM is expected to be finalised in August 2021, but it may be another 2-3 years before the model is fully ready for use. An important step to be performed is to show that the population dynamics, the colony structure and behaviour in the model is a good proxy for the real world. Then, further steps need to be taken before it can be used to predict pesticide effects (e.g. further harmonised data collection/generation). Finalisation of the key research projects mentioned above is crucial for supporting further ApisRAM calibration. The same actions would be required for the other bee models (i.e. bumble bees and solitary bees).

It is noted that the analysis presented in this document could be performed with new data and the bee models when they become available. A proper consideration of using ApisRAM (and any other bee models under development) can be fully performed once the models are finalised.

7. Reference tier (field studies) design in relation to the magnitude of acceptable effect

7.1. Preliminary estimation of the requirement for higher tier studies

The definition of the SPG, and particularly the selected "magnitude" of effect, has a direct impact on the requirements and feasibility of the reference tier testing (i.e. field studies).

To help risk managers make an informed decision, this section aims to provide a preliminary estimation of the higher tier requirements depending on the selected "magnitude" of acceptable effects. These requirements are expressed in terms of number of hives and field replication necessary.

These figures (reported in Table 4) need to be considered as preliminary, as some parameters used in this estimation (e.g. variability between colony sizes, level of type I and type II error as described in section 3.2.1) still need to be discussed and agreed by the Working Group dealing with the review of the EFSA bee guidance document. In addition, these estimations do not consider the increase in variability in time that colonies are likely to experience in field studies. The preliminary estimations in Table 4 show the number of fields and hives that would be needed to detect a certain threshold of acceptable effects. These preliminary estimations assume that five to eight hives are deployed per field, which represents a good coverage of the most common setups used in field studies.

Table 4: estimated total number of fields (i.e. treated + control fields) and bee hives needed in higher tier studies in order to detect different percentages of colony size reduction (i.e. magnitude of acceptable effect) with sufficient statistical power. This table assumes that 5 to 8 hives are monitored per field as an example.

	Thresholds of acceptable effect (i.e.% of colony size reduction) – SPG magnitude													
	1%	2%	3%	4%	5%	6%	7%	8%	9%	10%	12%	15%	20%	25%
8 hives/field														
Fields	944	234	104	58	38	26	20	14	12	10	6	4	2	2
Hives	7552	1872	832	464	304	208	160	112	96	80	48	24	16	16
7 hives/field														
Fields	1014	252	112	62	40	28	20	16	12	10	8	4	4	2
Hives	7098	1764	784	434	280	196	140	112	84	70	56	24	28	14
6 hives/field														
Fields	1108	276	122	68	44	30	22	18	14	12	8	6	4	2
Hives	6648	1656	732	408	264	180	132	108	84	72	48	36	24	12
5 hives/field														
Fields	1242	308	136	76	48	34	24	20	16	12	8	6	4	2
Hives	6210	1540	680	380	240	170	120	100	80	60	40	30	20	10

7.2. Example from available higher tier studies

EPPO (2010) offered recommendations for several types of experiments investigating pesticide effects on bees, including field studies. For these, it was specified that field replication is desirable, but "*often not feasible because of the requirements for separation*", which leads to a minimum requirement of one field for the treated group and one field for the control group. A minimum of four hives per field was also recommended. With this level of replication, and by using the same preliminary parametrisation used in the previous section, the minimum effect that can be detected as significant is around 25%, that is, when the mean colony size of the treatment group is at least 25% lower than that of the control group.

The most complex field studies ever evaluated by EFSA were considered in the context of the last review of the risk assessment for neonicotinoids[14] applied as seed treatment and granules.

Jaekel (2015) performed a study with one control and two treatment groups (two different substances applied), with one field per group and 6 hives per field. The experiment was replicated in 5 different countries (France, Germany, Hungary, UK, and Poland). Overall, 90 hives were deployed. In that case, the results were presented separately for each country. The same preliminary power analysis used before would estimate that the overall set-up could be able to detect as significant effects close to 10%. However, in that context, the testing on such a wide area would increase considerably the variability, thus decreasing the actual power.

Rolke et al. (2014) used 12 fields (6 for treatment and 6 for control) with 8 hives per field, with a total of 96 hives deployed for the experiment. Nevertheless, the study design presented some issues, as all treatment fields were clustered in one area and all control fields were clustered in another area, thus essentially reducing the replication to 1 treatment area and 1 control area. The post-hoc power analysis, i.e. using the variability observed in the study, carried out within the same project (with a different parametrisation than the one used above) revealed that the setup was able to detect as significant effects between 13.5 and 16.2%.

Rundlöf et al. (2015) used 16 fields (8 for the treatment and 8 for the control) with 6 hives per field, with a total of 96 hives deployed in the experiment. With the parametrisation used in section 7.1, this study design should be able to detect as significant effects close to 8%. The authors performed a power analysis as well, but with a different method and a different parametrisation than the one used in section 7.1. They concluded that their study was able to detect as significant effects of 19%. Extending the study one additional year using 10 fields (6 for the treatment and 4 for control) on the same farms, with 4 hives per field, totalling 40 hives, the authors reported that the setup over the two years resulted in a possibility to detect as significant effects below 5% (Osterman et al. 2019).

Finally, Woodcock et al. (2017) performed the study with the highest level of replication to date. They tested bees in 33 different fields: 11 for the control and 11 for each of the two treatments, as two substances were tested, scattered over three countries (Germany, Hungary, and UK). 6 hives per field were used, with a total of 198 hives deployed. With the parametrisation used for the other examples, this setup would be able to detect as significant effects close to 7%. However, it must be said that a more complex analysis performed by the same authors revealed that their study could detect only a considerably larger effect for the peak colony size with satisfactory power, partly because of the large variability among countries.

7.3. Considerations of the requirements of field studies in the EFSA bee guidance document

In the consultations held during the revision process of the bee guidance document, some stakeholders and MSs expressed concerns regarding the link between the SPG implemented in EFSA (2013) and its practical implementation in the risk assessment, particularly for the field studies. These concerns were related to the practicality and feasibility of detecting a magnitude of effects <7% of colony size reduction with a sufficient statistical power. In addition, concerns were also expressed in relation to the SPG consideration when evaluating field studies.

In particular, some MSs and stakeholders deemed the replication of field studies for addressing this requirement difficult to obtain considering that the foraging area of honeybees around the hive is large and the variation in vegetation and exposure to other pesticides may influence the results. It is difficult to ensure sufficiently controlled conditions for example to prevent cross-contamination.

Some MSs pointed out that to get 200 standardised colonies there is a need to start from 500 colonies (by leaving the weakest and strongest out) and only 2-3 apiculturists in MSs have such a number of colonies. Furthermore, standardisation will disappear in a hive after six weeks.

Overall, the statistical power and the requirements for field studies determined by EFSA (2013) for detecting the magnitude of 7% (e.g. number of replicates, field sites) was considered not feasible in realistic environments.

8. Concluding remarks for decision making process for risk managers

The present analysis provides scientific grounds to risk managers for the review of the SPG for honey bees, according to Step 3 of the EFSA method[2] and in particular as regards the quantification of the

magnitude of the effects. Feedback from risk managers is also required for the definition of an appropriate **temporal scale**.

In practice, the analysis of the background variability presented in this document should support risk managers in defining a threshold (or a set of thresholds based on geographical or seasonal variabilities) equivalent to a certain % reduction in colony size in a similar way as proposed by EFSA (2013).

When defining the threshold(s) of acceptable effects, risk managers should consider the following:

- **Colony size reductions of one third were already identified (see section 3.3) as a threshold potentially leading to the impairment of the colony viability.** The results of the simulations show that, in some circumstances (see Table 3), the distance between the mean size and the lower limit of the OR (either FOR or ROR) is close to or even higher than one third (i.e. 33%).
- **By using a more restricted OR the weaker colonies are left out, and therefore the threshold of acceptable effects is more conservative**. The results of the simulations show that only a few colonies are substantially weaker than the average colony size, even when no exposure to pesticide is present. By restricting the OR for threshold derivation, these weaker colonies will not be used as a reference for acceptable effects, resulting in a general higher protection level.
- **A restriction to the 50th percentile of the variability should not be considered for the threshold derivation**, since this would mean that, in many cases, only beneficial effects of pesticides are considered acceptable i.e. increase in colony size compared to the control.
- **The threshold of acceptable effect should be implementable in the reference tier**; since, currently, the reference tier is represented by the field studies, it should be realistically measurable in those studies (see section 7).

Additional considerations:

1) Why background variability can inform on the definition of the threshold(s) of acceptable effects

With approach #2, the magnitude dimension of the SPG will be based on a threshold of acceptable effect on colony size reduction identified within the range of the background variability, i.e. variability comparable to control replicates in experimental studies. Using the background variability means that, when evaluating a pesticide, the acceptable effect should not be larger than the variability of colonies not exposed to pesticides.

2) How reliable are the simulated variabilities?

- The variabilities simulated by the model were smaller than the median variability observed in the field studies, but still in the range of plausible values.
- The uncertainty of using simple landscape *vs* complex landscape was quantified. The analysis indicated that the variability in simple landscapes, such as the ones used in the current simulations, may be underestimated.
- The exclusion of *Varroa* from the simulations indicated that simulated variability may be underestimated
- The impact of other uncertainties is unknown.

An underestimation of the variability should be regarded as more conservative than an overestimation for the purpose of establishing an acceptable effect within the simulated background variability of colony size.

3) What are the consequences for risk assessment of the OR restrictions?

The narrower the range, the smaller the magnitude of the acceptable effect.

This means that there is a:

- higher level of conservativism

- lower impact on the provision of the ecosystem services
- lower trigger values for the tier risk assessment, thus higher number of substances that will not pass the risk assessment at the lower tiers
- higher requirements of field experiments (higher replication, higher costs, etc.) to reliably test whether the protection goal is met; therefore, practical limitations of the field studies should be considered.

4) What are the consequences for risk assessment if multiple thresholds are selected?

If risk managers select multiple thresholds of acceptable effects depending on the season/geographic variability, their implementation will require different risk assessments and the development and application of different criteria for the higher tier requirements. The consequence for the approval of the substance may be a highly demanding and non-harmonised risk assessment. This would also make it difficult to extrapolate a field study from one area/season to another.

5) What threshold should be selected in order to be implementable and measurable in field studies?

The selection of the threshold influences the required complexity of the experimental setup for field studies. EFSA is not in the position to express a definitive judgment of feasibility of different experimental setups, which mainly concerns costs and resources availability from the study sponsor. However, the examples of some of the most extensive studies that were recently evaluated (section 7), could be used as a benchmark. It is also important to note that the evaluation of complex field studies would require a high expertise and it is time- and resource-demanding.

References

Agatz A, Kuhl R, Miles M, Schad T, Preuss TG (2019). An Evaluation of the BEEHAVE Model Using Honey Bee Field Study Data: Insights and Recommendations. Environmental Toxicology and Chemistry, 38: 2535-2545.

Baude M, Kunin W, Boatman N, Conyers S, Davies N, Gillespie MAK, Morton RD, Smart SM, Memmott J (2016). Historical nectar assessment reveals the fall and rise of floral resources in Britain. Nature, 530, 85–88 (2016).

Becher MA, Grimm V, Thorbek P, Horn J, Kennedy PJ, Osborne JL (2014). BEEHAVE: A systems model of honeybee colony dynamics and foraging to explore multifactorial causes of colony failure. Journal of Applied Ecology, 51:470–482.

Becher MA, Grimm V, Knapp J, Horn J, Twiston-Davies G, Osborne JL (2016). BEESCOUT: A model of bee scouting behaviour and a software tool for characterizing nectar/pollen landscapes for BEEHAVE. Ecological Modelling, 340, 126-133.

Becher MA, Twiston-Davies G, Penny TD, Goulson D, Rotheray EL, Osborne JL, 2018. Bumble-BEEHAVE: a systems model for exploring multifactorial causes of bumblebee decline at individual, colony, population and community level. Journal of Applied Ecology, 55: 2790– 2801.

Bodenheimer FS, 1937. Studies in animal populations II. Seasonal population-trends in the honey-bee. The Quarterly Review of Biology 12, 406–425.

Bukovinszky T, Verheijen J. Zwerver S, Klop E, Biesmeijer JC, Wäckers F, Prins HH, Kleijn D, 2017. Exploring the relationships between landscape complexity, wild bee species richness and reproduction, and pollination services along a complexity gradient in the Netherlands. Biological Conservation, 214, 312-319.

Chauzat M-P, Cauquil L, Roy L, Franco S, Hendrikx P, Ribière-Chabert M (2013) Demographics of the European Apicultural Industry. PLoS ONE 8(11): e79018.

Cormont A, Siepel H, Clément J, Melman T, WallisDeVries M, Turnhout CV, Sparrius L, Reemer M, Biesmeijer JC, Berendse F, Snoo GR, 2016. Landscape complexity and farmland biodiversity: Evaluating the CAP target on natural elements. Journal for Nature Conservation, 30, 19-26.

Ebert GV, 1922. Zur Massenentwicklung der Bienenvoelker. Archiv für Biertenkunde. 4, 1-26, 37-38.

EFSA Panel on Plant Protection Products and their Residues (PPR), 2010. Scientific Opinion on the development of specific protection goal options for environmental risk assessment of pesticides, in particular in relation to the revision of the Guidance Documents on Aquatic and Terrestrial Ecotoxicology (SANCO/3268/2001 and SANCO/10329/2002). EFSA Journal, 8(10):1821, 55 pp.

EFSA Panel on Plant Protection Products and their Residues (PPR), 2012. Scientific Opinion on the science behind the development of a risk assessment of Plant Protection Products on bees (*Apis mellifera*, *Bombus* spp. And solitary bees). EFSA Journal, 10(5), 2668, 275 pp.

EFSA Panel on Plant Protection Products and their Residues (PPR), 2013. Guidance on tiered risk assessment for plant protection products for aquatic organisms in edge-of-field surface waters. EFSA Journal 2013;11(7):3290, 268 pp.

EFSA Panel on Plant Protection Products and their Residues (PPR), 2015. Statement on the suitability of the BEEHAVE model for its potential use in a regulatory context and for the risk assessment of multiple stressors in honeybees at the landscape level. EFSA Journal 2015; 13:4125.

EFSA Scientific Committee, 2016. Guidance to develop specific protection goals options for environmental risk assessment at EFSA, in relation to biodiversity and ecosystem services. EFSA Journal 2016;14(6):4499, 50 pp.

EFSA, 2013. EFSA Guidance Document on the risk assessment of plant protection products on bees (*Apis mellifera*, *Bombus* spp. and solitary bees). EFSA Journal, 11(7), 3295, 268 pp.

EFSA, 2015. EFSA Guidance Document for predicting environmental concentrations of active substances of plant protection products and transformation products of these active substances in soil. EFSA Journal 2015;13(4):4093, 102 pp.

EFSA, 2016. A mechanistic model to assess risks to honeybee colonies from exposure to pesticides under different scenarios of combined stressors and factors. EFSA supporting publication 2016:EN-1069. 116 pp.

EFSA, 2018. Evaluation of the data on clothianidin, imidacloprid and thiamethoxam for the updated risk assessment to bees for seed treatments and granules in the EU. EFSA supporting publication 2018:EN-1378. 31 pp.

EFSA, Ippolito A, del Aguila M, Aiassa E, Muñoz Guajardo I, Neri FM, Alvarez F, Mosbach-Schulz O, Szentes C, 2020a. Review of the evidence on bee background mortality. EFSA supporting publication 2020:EN-1880. 76 pp.

EFSA, Adriaanse P, Boivin A, Klein M, Jarvis N, Stemmer M, Fait G, Egsmose M, 2020b. Scientific report of EFSA on the 'repair action' of the FOCUS surface water scenarios. EFSA Journal 2020;18(6):6119, 301 pp.

European Commission, 2002. Guidance Document on Terrestrial Ecotoxicology under Council Directive 91/414/EEC (SANCO/10329/2002) rev.2 final, 17.10.2002, p.1 - 39.

European Commission, 2020. Honey market overview (Spring 2020). https://ec.europa.eu/info/food-farming-fisheries/animals-and-animal-products/animal-products/honey_en. [Accessed 20 July 2020].

EPPO, 2010. PP1/170(4) - Side-effects on honeybees. OEPP/EPPO Bulletin 40, 313–319.

Fijen TPM, Scheper JA, Boekelo B, Raemakers I, Kleijn D, 2019. Effects of landscape complexity on pollinators are moderated by pollinators' association with mass-flowering crops. Proceedings. Biological sciences, 286(1900).

FOCUS, 2000. FOCUS groundwater scenarios in the EU review of active substances. Report of the FOCUS Groundwater Scenarios Workgroup, EC Document Reference Sanco/321/2000 rev.2, 202pp

FOCUS, 2001. FOCUS Surface Water Scenarios in the EU Evaluation Process under 91/414/EEC. Report of the FOCUS Working Group on Surface Water Scenarios, EC Document Reference SANCO/4802/2001-rev.2. 245 pp.

Food and Agriculture Organization of the United Nation (FAO). FAOSTAT. http://www.fao.org/faostat/en/. [Accessed 18 June 2020].

Harbo JR, 1986. Effect of population size on brood production, worker survival and honey gain in colonies of honeybees. Journal of Apicultural Research 25(1): 22-29.

Hatjina F, Costa C, Büchler R, Uzunov A, Drazic M, Filipi J, Charistos L, Ruottinen L, Andonov S, Meixner MD, Bienkowska M, Dariusz G, Panasiuk B, Le Conte Y, Wilde J, Berg S, Bouga M, Dyrba W, Kiprijanovska H, Korpela S, Kryger P, Lodesani M, Pechhacker H, Petrov P, Kezic N, 2014. Population dynamics of European honey bee genotypes under different environmental conditions, Journal of Apicultural Research, 53:2, 233-247.

Jaekel KM,2015. Demonstration Farm Network – An approach to monitor health of honey bee colonies exposed to neonicotinoid seed-treated oilseed rape fields at pre-selected locations in France, Germany, Hungary, Poland and the United Kingdom 2014/15.

Leida B, Della Valle G, Piana L. I quaderni dell'apicoltore, 4. Flora Apistica. U.N.A.API – MIPAF.

Nowosad J, Stepinski TF, 2019. Information theory as a consistent framework for quantification and classification of landscape patterns. Landscape Ecology, 34, 2091–2101.

Ode Å., Hagerhall CM, Sang N, 2010. Analysing Visual Landscape Complexity: Theory and Application, Landscape Research, 35:1, 111-131.

Osterman J, Wintermantel D, Locke B, Jonsson O, Semberg E, Onorati P, Forsgren E, Rosenkranz P, Rahbek Pedersen T, Bommarco R, Smith HG, Rundlöf M, de Miranda J, 2019. Clothianidin seed-treatment has no detectable negative impact on honeybee colonies and their pathogens. Nature Communications 10: 692.

Persson AS, Olsson O, Rundlöf M, Smith HG (2010). Land use intensity and landscape complexity - Analysis of landscape characteristics in an agricultural region in Southern Sweden. Agriculture, Ecosystems and Environment, 136:169-176.

Rolke D, Persigehl M, Gruenewald B, Blenau W, 2014. Large-scale monitoring of long-term effects of Elado (10 g clothianidin & 2 g beta-cyfluthrin / kg seed) dressed oilseed rape on pollinating insects in Mecklenburg-Vorpommern, Germany: VII effects on honeybees (*Apis mellifera*).

Rundlöf M, Andersson GKS, Bommarco R, Fries I, Hederström V, Herbertsson L, Jonsson O, Klatt BK, Pedersen TR, Yourstone J, Smith HG, 2015. Seed coating with a neonicotinoid insecticide negatively affects wild bees. Nature, 521: 77-80.

Schmickl T, Crailsheim K, 2007. HoPoMo: A model of honeybee intracolonial population dynamics and resource management. Ecological modelling, 204, 219-245.

Timberlake TP, Vaughan IP, Memmott J (2019). Phenology of farmland floral resources reveals seasonal gaps in nectar availability for bumblebees. Journal of Applied Ecology, 56: 1585–1596.

Vicens and Bosch (2000). Weather-Dependent Pollinator Activity in an Apple Orchard, with Special Reference to *Osmia cornuta* and *Apis mellifera* (Hymenoptera: Megachilidae and Apidae). Environmental Entomology, 29(3): 413-420.

Wang B, Tian C, Sun J, 2019. Effects of landscape complexity and stand factors on arthropod communities in poplar forests. Ecology and Evolution, 9: 7143– 7156.

Woodcock BA, Bullock JM, Shore RF, Heard MS, Pereira MG, Redhead J, Ridding L, Dean H, Sleep D, Henrys P, Peyton J, Hulmes S, Hulmes L, Sárospataki M, Saure C, Edwards M, Genersch E, Knäbe S, Pywell RF (2017). Country-specific effects of neonicotinoid pesticides on honey bees and wild bees. Science, 356 (6345), 1393-1395.

Appendix A – Results of the simulation per regulatory zone

A.1. Results of the simulation for the Northern zone

A summary of the results for the simulations in the Northern zone scenarios is presented in Table A1 and Figure A1. The scenarios in this regulatory zone with lowest and the highest background variability are reported.

Table A1: percentage difference between the mean colony size and the lower limit of the OR for the Northern zone scenarios. The OR is presented as the whole variability range (i.e. the FOR) and as "restricted" variability ranges (RORs) to various extents.

Percentile of the variability as lower limit of the OR	% fraction of colonies retained in the OR	% difference between the mean colony size and the lower limit of the OR							
		High variability scenario (E4)				Low variability scenario (C4)			
		Full year	Spring	Summer	Autumn	Full year	Spring	Summer	Autumn
Whole range	100%	23.3%	25.0%	24.7%	22.6%	21.0%	25.4%	18.4%	20.7%
5th percentile	95%	14.7%	14.8%	16.0%	14.9%	11.8%	14.4%	9.3%	11.5%
10th percentile	90%	12.1%	12.0%	13.4%	12.3%	9.3%	11.0%	7.2%	9.3%
20th percentile	80%	8.6%	8.4%	9.6%	8.9%	6.2%	7.1%	4.5%	6.4%
30th percentile	70%	5.7%	5.4%	6.1%	6.2%	3.9%	4.4%	2.7%	4.1%
40th percentile	60%	2.9%	2.6%	2.8%	3.7%	1.9%	2.1%	1.1%	2.0%
50th percentile	50%	0.1%	-0.3%*	-0.5%*	1.0%	0.0%	0.0%	-0.2*%	0.1%

* Value > mean, should not be considered for threshold derivation

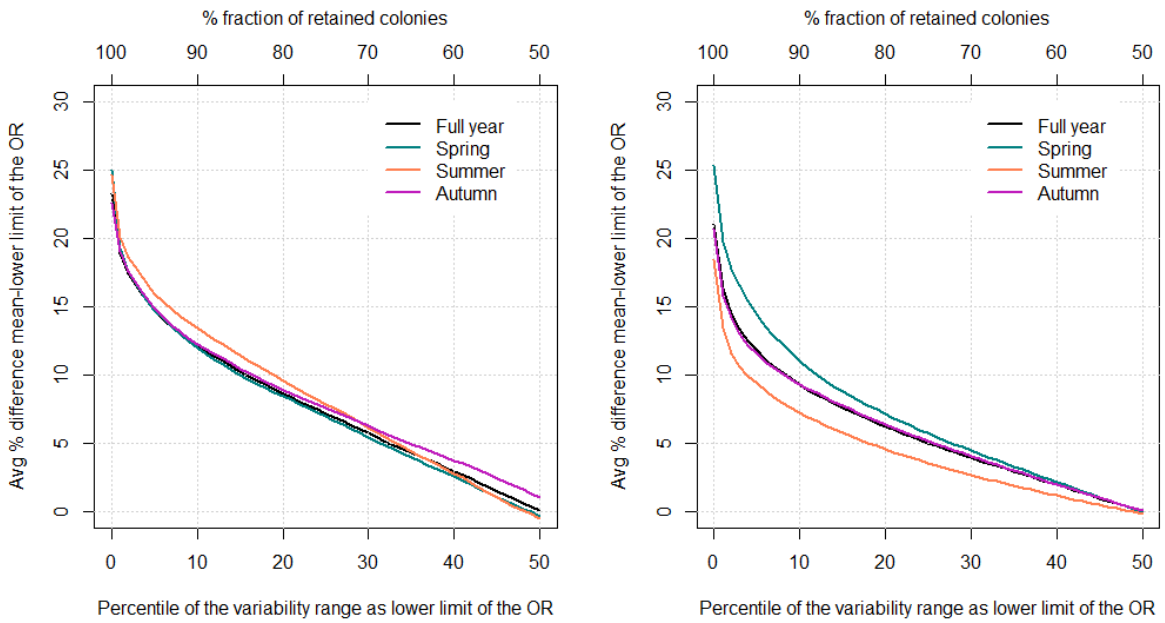

Figure A1: percentage fraction of colonies retained in the OR / percentile used as lower limit of the OR (upper/lower x-axis) vs. the related percentage difference between the mean and the lower limit of the OR (y-axis), averaged over one year and over single seasons, for the Northern zone scenarios. Left-hand side: E4 scenario (high variability); right-hand side: C4 scenario (low variability).

A.2. Results of the simulation for the Central zone

A summary of the results for the simulations in the central zone scenarios is presented in Table A2 and Figure A2. The scenarios in this regulatory zone with lowest and the highest background variability are reported.

Table A2: percentage difference between the mean colony size and the lower limit of the OR for the Central zone scenarios. The OR is presented as the whole variability (i.e. the FOR) and as "restricted" variability ranges (RORs) to various extents.

Percentile of the variability as lower limit of the OR	% fraction of colonies retained in the OR	% difference between the mean colony size and the lower limit of the OR							
		High variability scenario (E3)				Low variability scenario (D2)			
		Full year	Spring	Summer	Autumn	Full year	Spring	Summer	Autumn
Whole range	100%	31.1%	23.4%	24.2%	44.5%	20.4%	20.1%	12.8%	26.2%
5th percentile	95%	17.9%	13.3%	12.7%	27.2%	10.8%	10.9%	6.1%	13.8%
10th percentile	90%	13.2%	10.6%	9.6%	19.4%	8.7%	8.5%	4.8%	11.2%
20th percentile	80%	7.6%	7.0%	5.6%	10.3%	6.0%	5.7%	3.2%	7.8%
30th percentile	70%	3.6%	4.3%	1.5%	4.4%	3.9%	3.6%	2.0%	5.1%
40th percentile	60%	0.8%	1.9%	-0.2%*	0.3%	2.1%	1.8%	1.0%	2.8%
50th percentile	50%	-1.4%*	-0.1%*	-1.5%*	-2.8%*	0.2%	0.0%	0.0%	0.5%

* Value > mean, should not be considered for threshold derivation

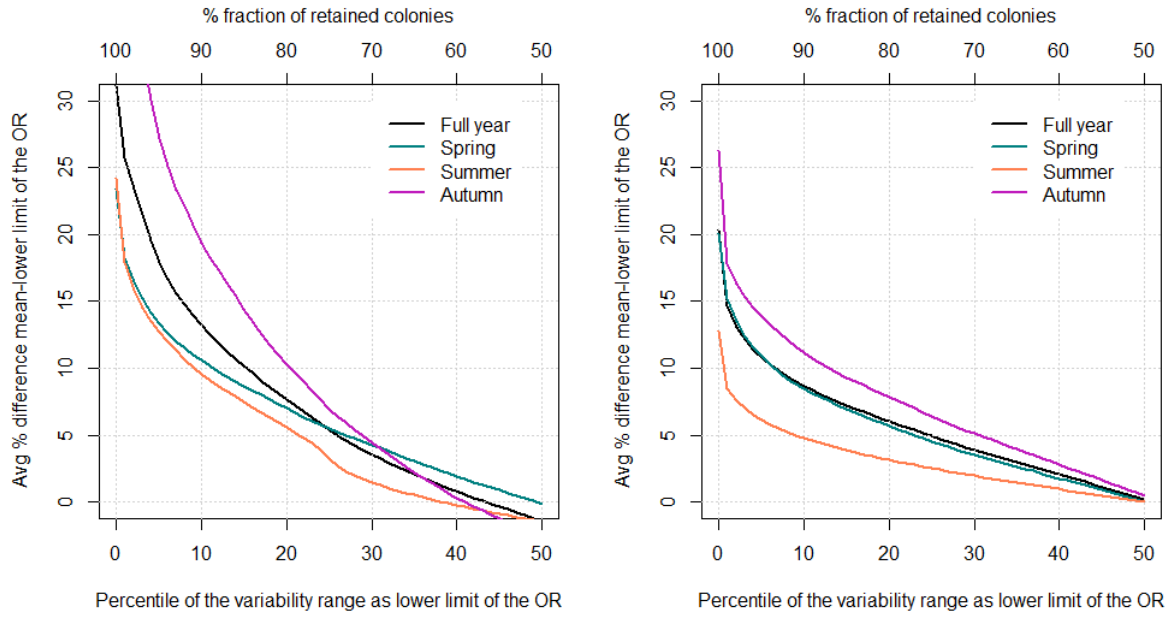

Figure A2: percentage fraction of colonies retained in the OR / percentile used as lower limit of the OR (upper/lower x-axis) vs. the related percentage difference between the mean and the lower limit of the OR (y-axis), averaged over one year and over single seasons, for the central zone scenarios. Left-hand side: E3 scenario (high variability); right-hand side: D2 scenario (low variability).

A.3. Results of the simulation for the Southern zone

A summary of the results for the simulations in the Southern zone scenarios is presented in Table A3 and Figure A3. The scenarios in this regulatory zone with lowest and the highest background variability are reported.

Table A3: percentage difference between the mean colony size and the lower limit of the OR for the Southern zone scenarios. The OR is presented as the whole variability (i.e. the FOR) and as "restricted" variability ranges (RORs) to various extents.

Percentile of the variability as lower limit of the OR	% fraction of colonies retained in the OR	% difference between the mean colony size and the lower limit of the OR							
		High variability scenario (A1)				Low variability scenario (C1)			
		Full year	Spring	Summer	Autumn	Full year	Spring	Summer	Autumn
Whole range	100%	28.8%	21.7%	47.1%	26.4%	20.3%	18.7%	19.4%	25.8%
5th percentile	95%	17.4%	12.4%	26.4%	17.8%	9.9%	9.5%	7.3%	12.6%
10th percentile	90%	13.3%	9.7%	18.3%	15.0%	7.3%	7.1%	5.6%	8.2%
20th percentile	80%	7.5%	6.1%	6.4%	11.1%	4.8%	4.5%	3.5%	5.2%
30th percentile	70%	4.3%	3.6%	1.7%	8.2%	3.0%	2.6%	2.0%	3.3%
40th percentile	60%	2.0%	1.5%	-1.0%*	5.7%	1.4%	1.1%	0.7%	1.5%
50th percentile	50%	-0.1%*	-0.3%*	-3.1%*	3.1%	-0.1%*	-0.2%*	-0.5%*	-0.2%*

* Value > mean, should not be considered for threshold derivation

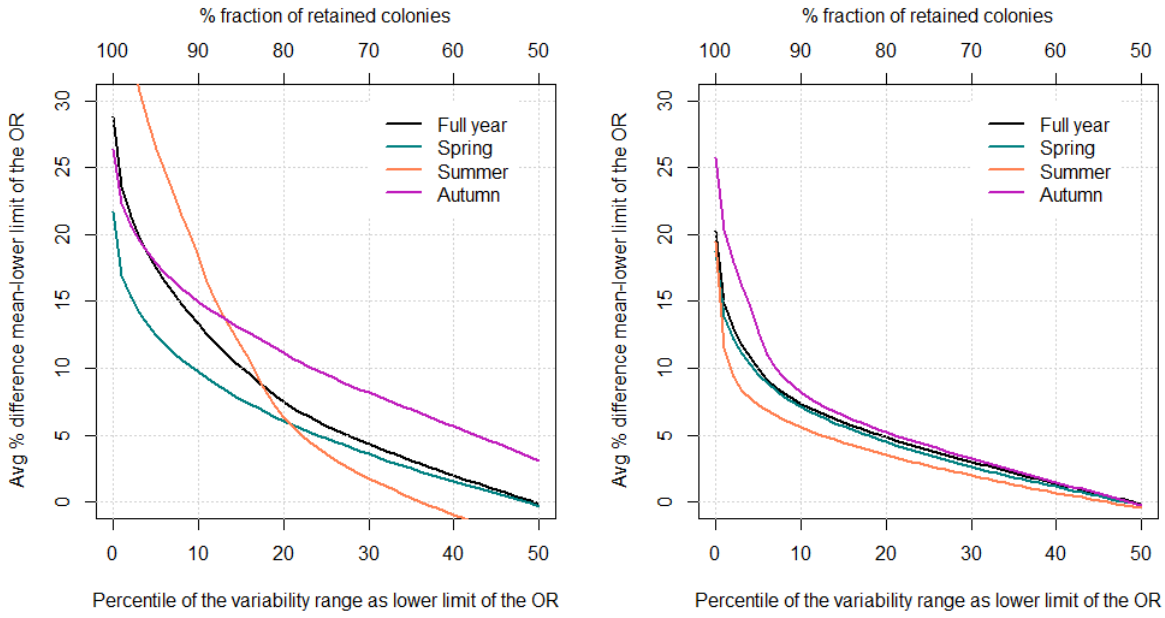

Figure A3: percentage fraction of colonies retained in the OR / percentile used as lower limit of the OR (upper/lower x-axis) vs. the related percentage difference between the mean and the lower limit of the OR (y-axis), averaged over one year and over single seasons, for the southern zone scenarios. Left-hand side: A1 scenario (high variability); right-hand side: C1 scenario (low variability).

Appendix B – Analysis of landscape complexity

B.1. Background of the issue

The default BEEHAVE landscape used in the analysis described in the main document consists of two flower patches, which are located at different distances from the hive, and have shifted phenology, but are in all other aspects (size, nectar provision, pollen provision, detection probability) identical.

Due to current limitation of data availability, a more realistic definition of landscape scenarios based on data was not possible. However, since the EFSA WG considered that the adopted simplification of the landscape was a crucial point, a separate analysis has been set up in order to explore the effect of landscape complexity on the final outcome (i.e. variability in colony size as simulated by the model).

Landscape complexity is defined in many ways in the literature, and even more operative ways for quantifying this concept are available. Nowosad & Stepinski (2019) highlighted that "*complexity is a concept defying a precise definition*" and they claim that "*there is no bona fide metric of landscape overall complexity*". Some authors have quantified landscape complexity by using the amount of natural and semi-natural habitats (see for example Bukovinszky et al., 2017; Cormont et al., 2016; Fijen et al., 2019). More structural attempts to identify the components of landscape complexity focused on the spatial arrangement of the landscape elements (see for example Wang et al., 2019 and Ode et al., 2010). Persson et al. (2010) analysed the relationship between landscape complexity and agricultural land use intensity, finding that they are separate factors, at least at smaller spatial scales.

In this Appendix it is investigated how changing various landscape parameters influences the output of BEEHAVE simulations, with a specific focus on the variability between equal replicate runs.

B.2. Dimensions of landscape complexity

Within the present work, EFSA have identified four dimensions which have the potential to regulate the foraging of the simulated bee hives. These are:

- Size heterogeneity of flower patches, i.e. the difference in size of food areas
- Patch fragmentation, i.e. the degree of scattering in the landscape of the different food patches
- Asynchrony of flowering, i.e. the degree of (lack of) overlap in the flowering period between patches
- Food level heterogeneity, i.e. the difference in food level provided by the different patches

The first two dimensions contribute to describe the spatial aspect of landscape complexity. The third concerns the temporal aspect, while the fourth captures flower source heterogeneity. For each of the four dimension, three different levels (low, medium, high) have been defined. All possible combinations of the three levels and four dimensions have been used for building 81 ($=3^4$) different landscapes (see fig B1 and B2). The overall size of the landscape (i.e. 3000m x 3000m) and the overall area of food patches in the landscape were kept constant, to avoid that food availability and not spatial or temporal aspects determine responses in the scenario simulations. In all landscapes, the food is provided by 9 "building block" patches (see fig. B1), which were changed in size and combined in space.

A description of the implementation of the different dimension is reported below, with the three levels always presented from the simplest to the most complex.

<u>Size heterogeneity of flower patches</u>

- Low: all patches have the same size (total food patch area/9).
- Medium: Patch size vary between 50 and 150%. The total food patch area is still the same.
- Large: Patch size vary between 10 and 200%. The total food patch area is still the same.

<u>Patch fragmentation</u>

- Low: all patches are directly neighboured to each other, building up 1 "field"

- Medium: patches are directly neighboured in group of 3. The resulting 3 "fields" are located in the upper, middle and lower part of the landscape.
- High: all patches are separated from each other, the 9 "fields" are distributed in the 9 quadrants of the landscape.

Asynchrony of flowering

- High synchronisation: all patches provide pollen and nectar exactly at the same time.
- Medium synchronisation: groups of three patches provide pollen and nectar at the same time, but each group has a shifted phenology compared to the others.
- Low synchronisation: each patch provides pollen and nectar at a different time.

Food level heterogeneity

- Low: all patches provide the same maximum level of pollen and nectar per square meter.
- Medium: the maximum level of pollen and nectar varies between each patch by a small extent (3 patches at 70%, 3 at 100%, 3 at 130%).
- Large: the maximum level of pollen and nectar varies between each patch by a large extent (between 25 and 300%).

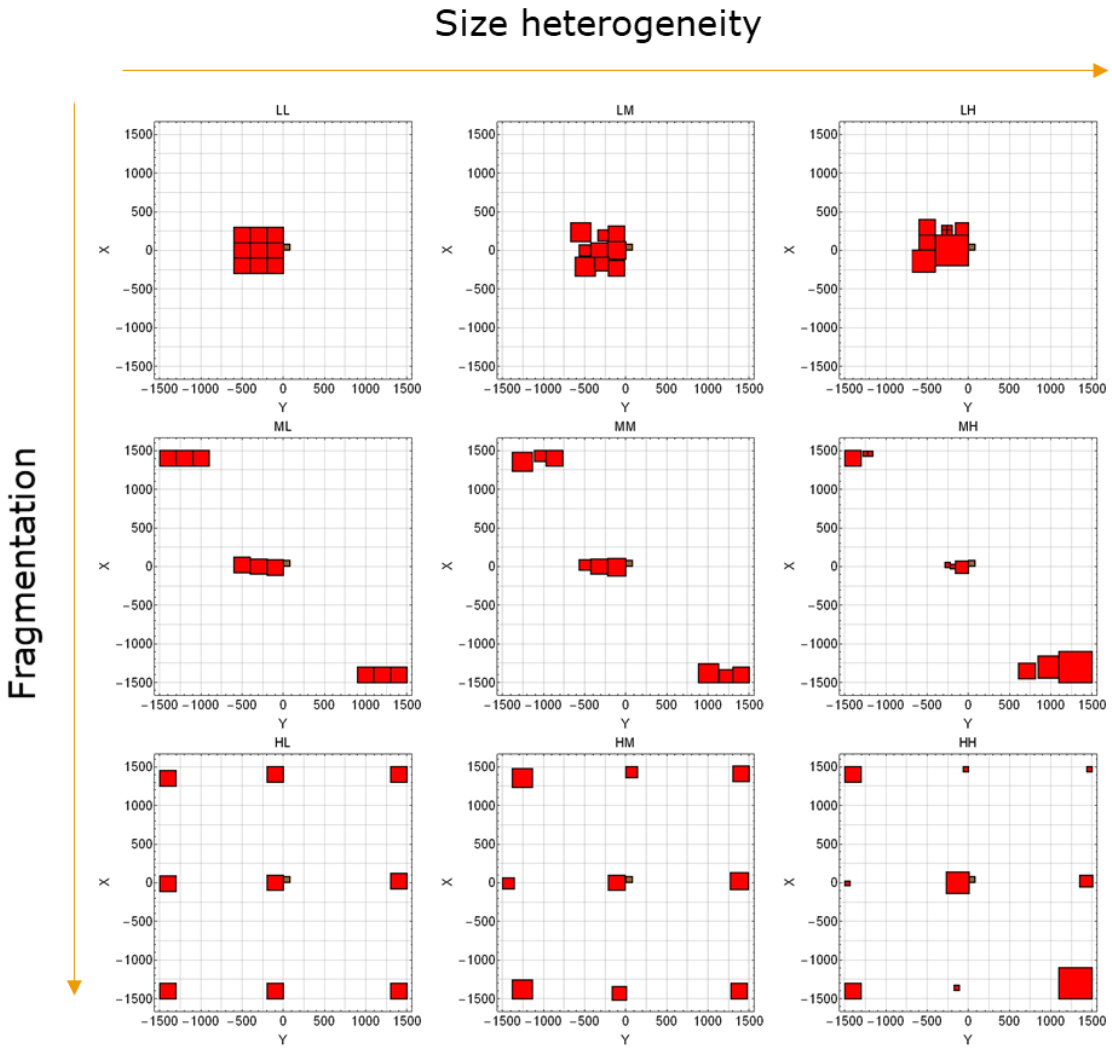

Figure B1: Spatial arrangement of the 9 food patches used in the analysis of landscape complexity. Size heterogeneity increases from left to right, while patch fragmentation increases from top to bottom. The hive is represented by the small square with the lower-left angle at the origin of the two axes (0,0).

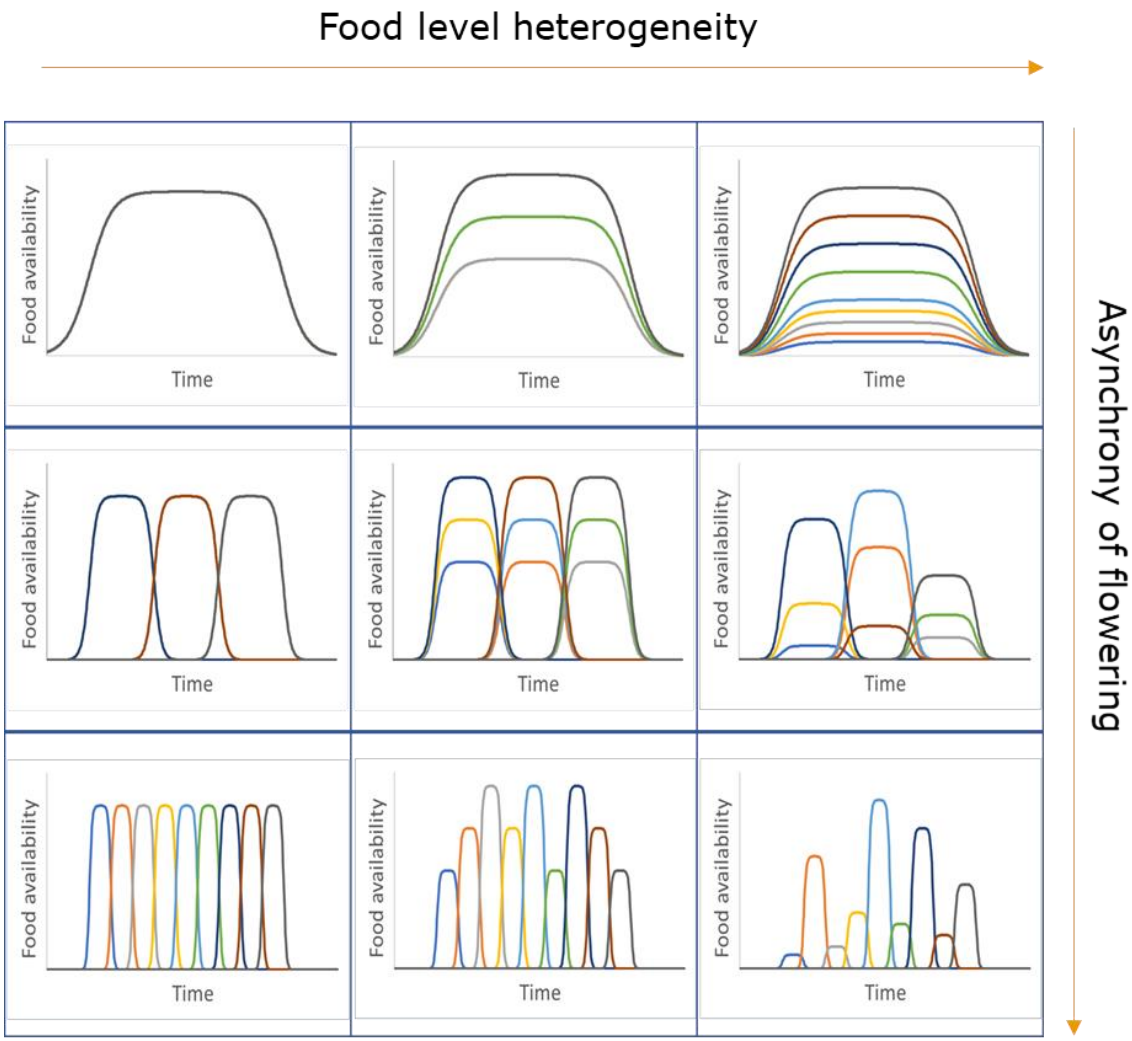

Figure B2: Temporal dynamics of food availability (pollen and nectar) in the 9 food patches used in the analysis of landscape complexity. The heterogeneity related to the maximum food level increases from left to right, while the asynchrony of flowering increases from top to bottom. When less than 9 curves are visible, this is because they are overlapping.

In addition, in order to explore the influence of landscape complexity under different climatic conditions, an entire latitudinal transect identified from the initially selected 20 locations was considered. Specifically, scenarios D1, D2, D3, D4, and D5 were used (see figure 1 in the main text). These locations span from Euboea (Greece) to Lapland (Finland).

The total flowering window in each landscape was adjusted on the basis of the foraging window (following the same approach used in the default 2-patches landscapes).

Overall, simulations were run in 405 different landscapes (81 landscape per scenario x 5 scenarios). 100 replicate runs were performed for each landscape.

B.3. Results of the analysis

The statistical analysis carried out on the output of the simulation confirmed that, in general, **landscape complexity increases variability in colony size**.

Asynchrony of flowering and fragmentation of patches were the main drivers of this process. Heterogeneity in flower patch size had some influence as well, while heterogeneity in food levels never showed a significant effect.

A common analysis was possible for scenarios D2, D3, D4, and D5 by using a linear (mixed) regression model. In order to make the outcome more understandable, the 10th percentile of the variability distribution was arbitrarily selected as the lower limit of the OR. The percentage difference between the mean colony size and this lower limit of the OR, averaged over the entire year, has been used as an indicator. This way the percentage differences reported hereafter are comparable with the values reported in table 3 of the main text and with tables A1-A3 in Appendix A.

This analysis showed that an increase of 1 level (e.g. from low to medium, or from medium to high) of flowering asynchrony, would cause the percentage difference to increase by 3.6%. Similarly, an increase of 1 level in flower patch fragmentation and size heterogeneity, would cause the percentage difference to increase by 2.2% and 0.6%, respectively. From the overall lowest level of complexity to the highest (considering all dimensions together), the model fitted to the simulation data predicts an increase in the percentage difference of 13.6%.

Scenario D1 showed a rather different outcome, with variabilities that achieved considerably higher values. The linear model fitted to the simulation data for D1 was able to satisfactorily describe the observed trend (R^2=0.76) by using only asynchrony of flowering, fragmentation of patches, and their interaction.

The presence of an interaction term causes the net effect of each dimension in isolation to be somehow harder to interpret. However, as only two explanatory variables are used in the final model, their joint influence can be easily visualised by means of a 3D plot (fig. B3).

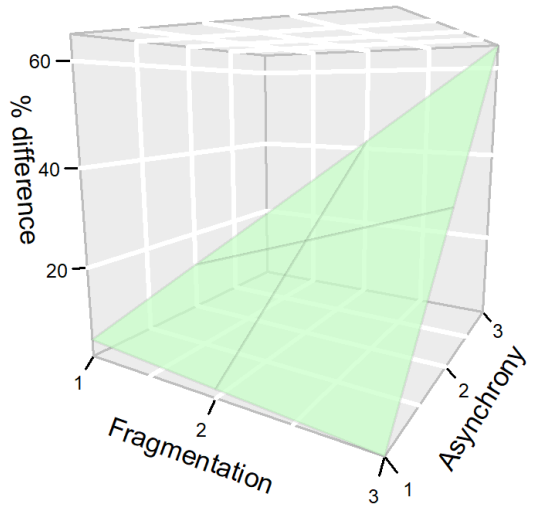

Figure B3: effect of patch fragmentation and asynchrony of flowering on the percentage difference between the mean and the lowest end of the OR (based on 10th percentile) for scenario D1. Relation as predicted by the linear model fitted to the simulation data.

The regression model indicates that the percentage difference may increase by more than 60% from the lowest to the highest level of the considered range of complexity in scenarios D1. This is obviously a much larger effect than the one observed for the other scenarios.

The reason for this larger effect is likely due to the scenario implementation performed in this analysis. In complex landscapes with high flowering asynchrony between patches, the flowering time in each patch is relatively short. This means that the bees have a relatively short time to discover the flowering patch, and not every patch will be discovered in every simulation run. This is further amplified in fragmented landscapes, as the detection probability decreases with the distance between the hive and the patch. In some replicate runs, bees discovered most food patches soon enough to use them. These runs are characterised by a rather continuous inflow of resources to the hive. On the contrary, in other replicate runs, several flowering events were missed. These can be characterised by gaps in the food inflow, with consequent problems in colony developments. Ultimately, the more complex the landscape, the more variability in the detection of food patches and food inflow, the more variability in colony size.

This is also confirmed by the correlation between average food (e.g. nectar) inflow variability and average colony size variability. This correlation is high for all scenarios (R^2 between 0.92 and 0.75).

Interestingly, the slope of this correlation follows a clear latitudinal gradient, suggesting that food inflow variability has a an even larger influence on colony size variability when the foraging season is longer. The larger effects seen in scenario D1 is very likely dependent on this last consideration, but perhaps not only. In scenario D1, missing a flowering patch would cause a rather long period of food deprivation, probably longer than larval development. So, if colonies were able to detect most patches "soon enough" they would develop without particular problems, otherwise they would struggle. This effect is mitigated in the other scenarios, as the shorter foraging window also caused a shorter time interval between flowering patches.

B.4. Conclusions

Overall, the outcome of the present analysis can be summarised in few rather simple conclusions:

- Landscape complexity produced an increase in colony strength variability
- The main drivers were asynchrony of flowering and patch fragmentation
- Landscape complexity increased colony size variability mainly via food inflow variability, which in turn was determined by different food patch detection probabilities
- The effect of food flow variability on colony size variability was stronger in scenarios with longer foraging periods

The quantitative effect of landscape complexity on the percentage difference between the mean and the lower limit of the OR (10^{th} percentile, in this case) was always significant, but somehow limited in scenarios D2, D3, D4, D5 (up to +13.6%). On the contrary, a large effect was seen in scenario D1 (up to + >60%).

The present exercise considered that the hive remained in the same location during the entire year, which is not always the case. In fact, many beekeepers move their colonies, especially in those situations when the food around the hive is scarce. In this sense the situations of higher landscape complexity used in this exercise may not be very frequent for many hives. However, this should be further checked by means of more in-depth analysis, which is not available for the time being.

Appendix C – Variability in risk assessment

In this Appendix a brief overview of the use of probability in the current environmental risk assessment of pesticides is presented.

In the past, risk managers have been asked several times to select thresholds within variability ranges and/or probability distributions (i.e. percentiles) for tuning the risk assessment procedures to the desired level of protection. So, in this respect, the current exercise does not represent a novelty.

Within the domain of the environmental risk assessment, examples of such selections are particularly abundant in the field of the exposure assessment, i.e. selecting the level of exposure to be used in the risk assessment.

The definition and related parameterisation of the FOCUS surface water scenarios aimed to cover at least "*a 90th percentile worst-case for surface water exposures resulting from agricultural pesticide use within the European Union*" (FOCUS, 2001). In other words, the estimated surface water concentrations used for the risk assessment should be higher than 90% of possible situations occurring in time and space (EFSA et al., 2020b).

Similarly, weather and soil characteristics used in the FOCUS ground water scenarios are combined to result in an overall 90th percentile vulnerability in terms of leaching (FOCUS, 2000). The same 90th percentile goal is also recommended for soil, although this will be applied specifically for each regulatory zone (EFSA, 2015).

The risk assessment for bees in EFSA (2013) is based on an exposure estimation that should cover the 90% (i.e. the 90th percentile) of the hives placed in the vicinity of the treated fields.

The use of probability and/or variability thresholds in risk assessment is however not limited to exposure. Species sensitivity distributions (SSD) are often used to derive HC_5 (hazardous concentration for 5% of the species, i.e. the 5th percentile), or analogous hazard estimates (e.g. HR_5, HP_5). This means that the selected toxicological threshold is protective for 95% of the species. For example, the current risk assessment for non-target plants (European Commission, 2002) considers the SSD as a tool to identify an application rate that would protect (i.e. causing <50% effect) 95% of the species being potentially exposed, leaving out 5% of those. Since no assessment factor is used in such risk assessment, this percentage could immediately be interpreted as the current protection goal, although this was never explicitly agreed.

SSD-derived hazard estimates are also used for assessing risk to aquatic organisms (EFSA PPR panel, 2013). However, the presence of assessment factors complicates the relations between the threshold (5%, i.e. the 5th percentile) of the sensitivity distribution used in the risk assessment and the final object of protection.

All previous examples deal with situations when one of the two main dimensions of the risk assessment (i.e. either the exposure or the hazard) is known to be variable: across space and time, across species, etc. In order to make the risk assessment operational, several times in the past there has been an explicit decision to neglect part of the variability distribution, generally the most "extreme" part. In this respect, the present task for the risk managers does not represent a novelty.

Nevertheless, the present task presents two fundamental differences compared to similar decisions made in the past.

1. The selection of the percentile, unlike all other situations, has an upper bound at the mean of the variability distribution (see figure 1) i.e. often close to the 50th percentile[15]. Selecting a percentile higher than the mean as lower limit of the OR would indeed be nonsensical, as any treatment would need to produce a positive effect on the mean colony size to be considered acceptable.
2. Second, in all examples presented in this section, "pushing" the threshold towards most extreme values (in particular very high percentiles for the exposure assessment goals and very

[15] Most variability distributions are not heavily skewed (i.e. more or less symmetric), thus the mean and the median are not too dissimilar.

low percentiles for the SSD) would translate into a more conservative risk assessment. On the contrary, in the present task, the selection of very low percentiles as lower limit of the OR would translate in less conservative risk assessments as this would allow for a higher acceptable effect.

By putting together the two points above it derives that **the conservativeness of risk assessment will increase with percentiles that get closer to the mean** (i.e. likely closer to 50th percentile).